On the Stationarity of Multivariate Time Series for Correlation-Based Data Analysis

Kiyoung Yang and Cyrus Shahabi
Computer Science Department
University of Southern California
Los Angeles, CA 90089-0781
[kiyoungy,shahabi]@usc.edu

Abstract

Multivariate time series (MTS) data sets are common in various multimedia, medical and financial application domains. These applications perform several data-analysis operations on large number of MTS data sets such as similarity searches, feature-subset-selection, clustering and classifications. Correlation-based techniques, such as Principal Component Analysis (PCA), have proven to improve the efficiency of many of the above-mentioned data-analysis operations on MTS, which implies that the correlation coefficients concisely represent the original MTS data. However, if the statistical properties (e.g., variance) of MTS data change over time dimension, i.e., MTS data is non-stationary, the correlation coefficients are not stable. In this paper, we propose to utilize the stationarity of the MTS data sets, in order to represent the original MTS data more stably, as well as concisely with the correlation coefficients. That is, before performing any correlation-based data analysis, we first executes the stationarity test to decide whether the MTS data is stationary or not, i.e., whether the correlation is stable or not. Subsequently, for a non-stationary MTS data set, we difference it to render the data set stationary. Even though our approach is general, to focus the discussion we describe our approach within the context of our previously proposed technique for MTS similarity search. In order to show the validity of our approach, we performed several experiments on four real-world data sets. The results show that the performance of our similarity search technique have significantly improved in terms of precision/recall.

1 INTRODUCTION

A time series is a series of observations, $x_i(t); [i = 1, \cdots, n; t = 1, \cdots, m]$, made sequentially through time where i indexes the measurements made at each time point t [17]. It is called a univariate time series (UTS) when n is equal to 1, and a multivariate time series (MTS) when n is equal to, or greater than 2. A UTS data is usually represented in a vector of size m, while each MTS data is typically stored in an $m \times n$ matrix, where m is the number of observations and n is the number of variables.

An MTS data is typically very high dimensional. For example, an MTS data from one of the data sets used in the experiments in Section 3 contains 3000 observations with 64 variables. If a traditional distance metric for similarity search, e.g., Euclidean Distance, is to be utilized, this MTS data would be considered as a 192000 (3000×64) dimensional data. 192000 dimensional data would be overwhelming not only for the distance metric, but also for indexing techniques. To the best of our knowledge, there has been no attempt to index data sets with more than 100000 dimensions/features[1]. Hence, instead of using this high dimensional MTS data set as is, a number of techniques have been proposed to represent the MTS data set concisely for data mining processes, such as classification and similarity search [14, 13, 16, 18, 21]. For example, in [13], an MTS data that contains an EEG signal with 39 channels is decomposed into multiple UTSs. Each UTS is subsequently transformed into 3 autoregressive (AR) coefficients. Hence, each MTS data is represented with 117 (3×39) features, after which feature subset selection is performed.

For MTS analysis, e.g., similarity search and feature subset selection, it has been empirically shown that the correlation information among the variables plays an important role, and the correlation-based techniques, such as PCA, perform well [16, 14, 18]. In [20], each MTS data is represented with the upper triangle elements of the correlation coefficient matrix. The performance of feature subset selection using the correlation coefficients is shown to out-

[1] In [1], the authors employed MVP-tree to index 65536 dimensional gray-level MRI images.

Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
1550-4786/05 $20.00 © 2005 IEEE

perform the one using the AR coefficients which does not consider the correlation information among the variables. In [16], Principal Component Analysis (PCA) Similarity Factor (S_{PCA}) [11] is employed for the similarity measure between two MTS data. That is, in order to compute the similarity between two MTS data, they first obtain the correlation coefficient matries[2] of the two MTS data, and then decompose them via Singular Value Decomposition (SVD) to obtain the principal components. Consequently, they measure how similar the corresponding principal components (PCs) from the two MTS data are.

To recapitulate, the correlation-based techniques utilize the correlation coefficients to represent the original MTS data for data mining tasks. The good performance of correlation-based techniques hence implies that the correlation coefficients concisely represents the original MTS data in a dimension-reduced form. Note that, given an MTS data $\mathbf{A}$ of size $m \times n$, it is typically the case that $m \gg n$. For example, an MTS data from one of the data sets used in Section 3 contains 3000 observations, while there are only 64 variables. In general, the size of a correlation coefficient matrix for $\mathbf{A}$, i.e., $n \times n$, is much smaller than that of $\mathbf{A}$.

However, if the properties of time series change over time, i.e., the time series is not *stationary*, then the correlation coefficients are not *stable*. For example, assume that we are given an MTS data $x_i(t)$ of size $m \times n$, i.e., where $1 \leq i \leq n$ and $1 \leq t \leq m$. From $x_i(t)$, consider two matrices of size $(m-1) \times n$, i.e., $x_i^1(t)$ where $1 \leq t \leq m-1$ and $x_i^2(t)$ where $2 \leq t \leq m$. If $x_i(t)$ is *non-stationary*, then $Corr(x_i^1(t))$ and $Corr(x_i^2(t))$ would be statistically different. This example implies that for non-stationary MTS data, the correlation coefficients change *statistically significantly* depending on just one observation out of, e.g., 3000 observations, which is not the case for stationary MTS data.

In this paper, we propose to utilize the *stationarity*[3] of time series in order to better represent the MTS data with the correlation coefficients. Intuitively, a time series is defined to be *stationary* if the statistical properties of the time series, e.g., the mean and the correlation coefficients, do not change over time. Hence, if an MTS data is stationary, the correlation information of the MTS data does not change over time, which would make the correlation based representations of the original MTS data more *stable*. Therefore, we firstly test the stationarity of each MTS data in the database. Subsequently, we determine the stationarity of the MTS data set based on the majority of stationarities of MTS data in the data set. If the MTS data set turns out to be non-stationary, we *stationarize* all the MTS data in the data set before we perform correlation-based data analysis.

In order to evaluate the effectiveness of the proposed approach, we conducted several experiments on four real-world data sets. Even though our approach is general, to focus the discussion we describe our approach within the context of our previously proposed technique for MTS similarity search called Eros [18]. The performances depending on the stationarity of the data set have been compared in terms of the precision/recall using Eros. The results show that the performance improves up to 24% in precision/recall.

2 THE PROPOSED APPROACH

Algorithm 1 Determine the stationarity of an MTS data set

Require: MTS data set, N {the number of data in the data set}, n {the number of variables in an MTS data}

1: **for** i = 1 to N **do**
2: $res \leftarrow$ Johansen's test on the ith MTS data;
3: $\gamma \leftarrow$ extract the number of co-integrating relationships from res;
4: **if** $\gamma = n$ **then**
5: $H(i) \leftarrow 0$; {stationary}
6: **else**
7: $H(i) \leftarrow 1$; {non-stationary}
8: **end if**
9: **end for**
10: **if** $sum(H) > ceil(N/2)$ **then**
11: Stationarize the given MTS data set;
12: **end if**

Before performing any correlation-based data analysis for multivariate time series, we propose to render the data set as stationary, if necessary, as in Algorithm 1. That is, we firstly determine the stationarity for all the MTS data in the data set by performing the Johansen's Co-integration test [8], for implementation of which the Econometrics Toolbox [4] is employed. Note that we do not utilize the Augmented Dickey Fuller (ADF) test [3] for each UTS, since we are not interested in the stationarity of each UTS in an MTS data. Besides, as in [5], ADF test requires a strategy that deals with three cases to appropriately apply the ADF test to an UTS, which seems to be rather cumbersome. Moreover, it has been shown that ADF test has low power, failing to reject the unit root hypothesis in many cases [15, 2].

As in Line 3 of Algorithm 1, we extract the number of co-integrating relationships, γ, where $1 \leq \gamma \leq n$. As long as γ is equal to the number of variables of an MTS data, n, we consider the MTS data as stationary (Lines 4∼8). The stationarity of an MTS data set is subsequently determined by the majority of the stationarities of the MTS data. The

[2]PCA may employ either the correlation coefficient matrix or the covariance matrix for a given MTS data. We would assume that the correlation coefficient matrix is utilized for PCA, since correlation-based analysis is more frequently encountered [9].

[3]For details on the stationarity and how to test the stationarity of a time series, please refer to [8, 3, 4, 6].

[4]http://www.spatial-econometrics.com/

Table 1. Summary of data sets used in the experiments

	AUSLAN	BCI	BCI MPI	EEG
# of variables	22	64	39	64
average length	60	3000	1280	256
# of labels	95	2	2	2
# of MTS items per label	27	139	1000	6980/3908
total # of MTS items	2565	278	2000	10888

Table 2. Stationarity Test Results

Data set	Co-integration Test	
	# of stationary data	# of non-stationary data
AUSLAN	1311	792
BCI	233	45
BCI MPI	309	1691
EEG	954	9934

$sum(H)$ in Line 10 yields the number of non-stationary data in the data set. Hence, if the number of non-stationary MTS items is greater than half of the total number of MTS items, the data set is determined to be non-stationary, and each MTS item is *first-order differenced* into a stationary item. Intuitively, if we make sure that the data set is stationary, the original MTS data can be more *stably*, as well as *concisely*, represented with correlation coefficients.

3 PERFORMANCE EVALUATION

In order to evaluate the effectiveness of our proposed approach, we compared the impact of stationarity on MTS data sets within he context of our previously proposed similarity measure called Eros [18]. For a stationary data set, we compare the performances of Eros without differencing to those with differencing, and see how much they improve, and vice versa for a non-stationary data set. The Johansen's Co-integration test has been performed with a significance level of 5%. The experiments have been conducted on four different real-world data sets, i.e., AUSLAN [10], BCI [12], BCI MPI and EEG, which are all labeled MTS data sets whose labels are given. Table 1 shows the summary of the data sets used in the experiments.

We performed modified leave-one-out kNN search as in [18]. For simplicity, we chose 10 for $maxr$. Recall that each data set used in the experiments has more than 10 similar items per label as shown in Table 1. For example, AUSLAN has 95 labels and with 27 items per label. The recall-precision graph [7] is then plotted, which has been frequently used to measure the performance of Content Based Image Retrieval (CBIR) systems as well as Information Retrieval (IR) systems. The *mean* aggregating function on the raw eigenvalues has been used for the weight vector w of *Eros*.

3.1 RESULTS

Table 2 summarizes the results of Johansen's Co-integration test on the four data sets. AUSLAN and BCI are determined to be stationary, while BCI MPI and EEG are non-stationary, i.e., less than half of the MTS items are stationary. Figure 1 depicts the precision/recall using Eros. For all the data sets, the better performance corresponds to the stationarities of the data sets. That is, if the data set is stationary, the better performance is obtained without differencing, and if the data set is non-stationary, the better performance is achieved with differencing.

The performance improvements between with and without differencing are up to 24% in terms of precision/recall. Hence, our proposed approach to utilizing stationarity is well justified. For more detailed experiments and discussion, please refer to [19].

4 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to render the given MTS data set as stationary, if necessary, before performing correlation-based data analysis, such as PCA. Based on the stationarity, the correlation coefficients represent the original MTS data more *stably* as well as concisely. Empirically, we have shown that if the given data set is non-stationary, making the data set stationary improves the performances of the correlation-based data analysis in terms of precision/recall.

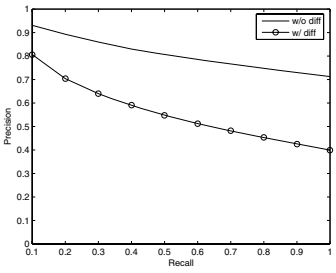
We intend to extend this technique to the stream of data where the determination of the stationarity as well as the subsequent data analysis processes, such as, feature subset selection, can be performed incrementally adjusting itself based on the observations collected thus far.

Acknowledgement

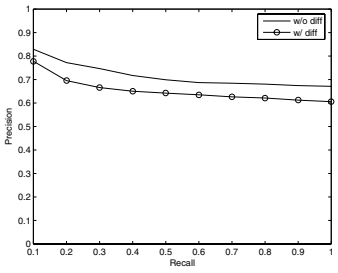
This research has been funded in part by NSF grants EEC-9529152 (IMSC ERC), IIS-0238560 (PECASE) and IIS-0307908, and unrestricted cash gifts from Microsoft. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

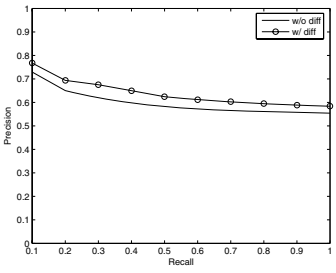
[1] T. Bozkaya and M. Ozsoyoglu. Indexing large metric spaces for similarity search queries. *ACM TODS*, 24(3), 1999.



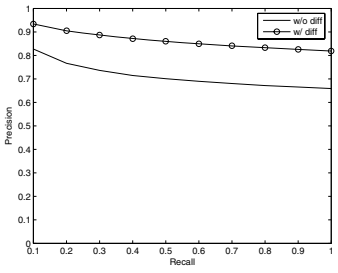
(a) AUSLAN (stationary)



(b) BCI (stationary)



(c) BCI MPI (non-stationary)



(d) EEG (non-stationary)

Figure 1. Precision/Recall using Eros

[2] D. N. DeJong, J. C. Nankervis, N. E. Savin, and C. H. Whiteman. Integration versus trend stationary in time series. *Econometrica*, 60(2):423–433, March 1992.

[3] D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431, June 1979.

[4] D. A. Dickey, D. W. Jansen, and D. L. Thornton. A primer on cointegration with an application to money and income. *Federal Reserve Bulletin, Federal Reserve Bank of St. Louis*, pages 58–78, March/April 1991.

[5] J. Elder and P. E. Kennedy. Testing for unit roots: What should students be taught? *Journal of Economic Education*, 31(2):137–146, 2001.

[6] R. F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, March 1987.

[7] W. B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992.

[8] S. Johansen. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, 1995.

[9] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[10] M. W. Kadous. *Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series*. PhD thesis, University of New South Wales, 2002.

[11] W. Krzanowski. Between-groups comparison of principal components. *JASA*, 74(367), 1979.

[12] T. N. Lal, T. Hinterberger, G. Widman, M. Schröder, N. J. Hill, W. Rosenstiel, C. E. Elger, B. Schölkopf, and N. Birbaumer. Methods towards invasive human brain computer interfaces. In *Advances in Neural Information Processing Systems 17*, pages 737–744. Cambridge, MA, 2005.

[13] T. N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf. Support vector channel selection in BCI. *IEEE Trans. Biomed. Eng.*, 51(6), June 2004.

[14] C. Li, P. Zhai, S.-Q. Zheng, and B. Prabhakaran. Segmentation and recognition of multi-attribute motion sequences. In *ACM MM '04*, pages 836–843, New York, NY, USA, 2004.

[15] P. Perron. The great crash, the oil price shock, and the unit root hypothesis. *Econometrica*, 57(6):1361–1401, 1989.

[16] A. Singhal and D. Seborg. Clustering of multivariate time-series data. In *Proc. of the American Control Conference*, volume 5, 2002.

[17] A. Tucker, S. Swift, and X. Liu. Variable grouping in multivariate time series via correlation. *IEEE Trans. Syst., Man, Cybern. B*, 31(2):235–245, 2001.

[18] K. Yang and C. Shahabi. A PCA-based similarity measure for multivariate time series. In *MMDB '04*, pages 65–74, Washington, DC, USA, 2004.

[19] K. Yang and C. Shahabi. On the stationarity of multivariate time series for correlation-based data analysis. Technical report, University of Southern California, 2005.

[20] K. Yang, H. Yoon, and C. Shahabi. A supervised feature subset selection technique for multivariate time series. In *FSDM*, Newport Beach, CA, April 2005.

[21] H. Yoon, K. Yang, and C. Shahabi. Feature subset selection and feature ranking for multivariate time series. *IEEE Trans. Knowledge Data Eng.*, 17(9), September 2005.