A KST framework for Correlation Network Construction from Time series signals

Jin-Peng Qi^a, Quan Gu^{a,b}, Ying Zhu^c, Ping Zhang^{*d}

^a College of Information Science & Technology, Donghua University, Shanghai 201620, P.R. China

^b MRC-University of Glasgow Centre for Virus Research, Glasgow, G61,1QH, Scotland, UK

^c Hunter New England Health, Royal North Shore Hospital, NEW South Wales, Australia

^{*d} Menzies Health Institute Queensland, Griffith University, Australia

p.zhang@griffith.edu.au

ABSTRACT

A KST (Kolmogorov–Smirnov test and T statistic) method is proposed for construction of a correlation network based on the fluctuation of each time series within the multivariate time signals. In this method, each time series is divided equally into the multiple segments, and the maximal data fluctuation in each segment is calculated by our KST change detection procedure. Connections between each time series are derived from the data fluctuation matrix, and are used for constructing the fluctuation correlation network (FCN). The method was tested with synthetic simulations and the result was compared with those from using KS or T only for detection of data fluctuation. The novelty of this study is that the correlation analyses was based on the data fluctuation in each segment of each time series rather than on the original time signals. It would be more meaningful for many real world applications and for analysis of large-scale time signals where prior knowledge is uncertain.

Keywords: Time series, Abrupt change, Data fluctuation, Correlation network.

1. INTRODUCTION

Many systems of scientific and societal interest are composed of a large number of interacting elements, and can be described with correlation networks, where network nodes represent the elements in a given system and network ties represent interactions between the elements [1, 2]. Correlation networks have been studied in many financial and biological contexts, such as stock market price, mobile communication and gene expression data [3-5]. Multivariate time series analysis has also attracted special attention due to its practical and theoretical importance of complex systems in physics, physiology, biology and society etc [6, 7]. Network theory has shown to be a powerful tool for revealing information embedded in multivariate or multi-channel time series [5, 8], and it has led to an avalanche of findings of non-trivial features in time series, including the long-range correlations, the scale invariance and so on [9, 10]. The structure of a correlation network built based on the multivariate time series might be associated with some functionality for example in biological application, and the change of network might be associated with the change of some corresponding functionality [11]. With medical electroencephalogram (EEG) as an example, the correlation network built on the signals from different locations in a brain may reveal the functional connection between the different parts of the brain. An unexpected change of the network structure may help analyze the condition change of a patient. Authors in [12] combine multivariate time series analysis and graph theory to detect synchronization and causality among certain ecological variables and to represent significant interactions via network projections. Other researchers, such as [13, 14], also studied the correlation network of time series to detect the functional change associated with the time series.

In many cases, the correlation networks are extremely complex and network structures are continuously evolving through various heterogeneous interactions in time [9, 15]. In [16], authors reviewed their systematic works on autocorrelations in financial time series data, from the perspective of complex networks. Researchers have put lots of effort on improving the algorithms for network construction. The correlation networks created for different applications were all based on the original signals. However, in reality most of the large-scale time signals, non-stationary and noisy

[17], and only a small part of total amount have vital contribution for correlation network construction. Therefore the interpretation of these networks might not be able to represent the real application. On the other hand, the correlation between the fluctuations of time signals may have better link to the connection of associated functionalities. In this research we propose a method to build a correlation network based on the time signal fluctuations in multivariate time series. A significant fluctuation in a time series can be considered as a change point. To construct a meaningful correlation network based on the changes of the signals in a multivariate time series, the change at the time point has to be detected accurately beforehand. In fact, there are many studies focuses on time point change detection algorithms. A review of these algorithms can be found in [18]. In our previous study on time point change detection [19-21], we analyzed how KS and T could be used for calculating the fluctuations of a time series. In this study, we took the advantages of KS and T statistic for detection of the maximum fluctuations in the segments of a time series and built a fluctuation correlation network (FCN) based on the fluctuation data matrices that was created from a combined method with KS and T statistics (KST). We also compared the result from the KST method with using only KS or T. The rest of the paper is organized as follows. Section 2 gives the mathematical framework of the proposed KST method. Section 3 presents the implementation and experiments on the synthetic time signals. Section 4 concludes the paper and suggests future work.

2. METHOD

To construct a FNC, it is important to study and figure out the interactions between each pair of multivariate time series [22]. Each time series can be segmented into multiple sections, and a correlation matrix can be created for a multivariate time series based on the signal fluctuation in each section/segment of each time series. In this study, we used the KS and T statistic to detect the fluctuation in each section of the time series and constructed a correlation network that represent the correlations between each time series within a multivariate time series.

2.1 Overview of the KST Framework

Generally, KS and T can be used for change or fluctuation detection from a time series. However, the KS-based method is insensitive to the data fluctuation near the left and right boundaries, especially when the sample size is small. Meanwhile, the T-based method is also inefficient for the fluctuation near the middle boundaries, especially when the sample size is bigger [23, 24]. To resolve these problems, we combined KS and T and proposed a KST framework for the fluctuation matrix construction.

First, for each time series $S_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,N}\}$ in multivariable time signals $\{S_1, S_2, \dots, S_m\}, 1 \leq i \leq m$, we divide S_i equally into n segments and get a segment vector $Ms_i = \{S_{i,1}, S_{i,2}, \dots, S_{i,n}\}$, $n = \log_2 N$, and then obtain a segment matrix Ms as,

$$Ms = \begin{bmatrix} Ms_1 \\ Ms_2 \\ \dots \\ Ms_m \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,n} \\ S_{2,1} & S_{2,2} & \dots & S_{2,n} \\ \dots & \dots & \dots & \dots \\ S_{m,1} & S_{m,2} & \dots & S_{m,n} \end{bmatrix}, \quad (1)$$

Suppose the data fluctuation in each element of Ms is measured respectively by KS and T, and the KS-based fluctuation matrix termed Msf_{ks} and the T-based fluctuation matrix termed Msf_t are denoted as,

$$Msf_{ks} = \begin{bmatrix} Msf_{ks1} \\ Msf_{ks2} \\ \dots \\ Msf_{ksm} \end{bmatrix} = \begin{bmatrix} ksF_{1,1} & ksF_{1,2} & \dots & ksF_{1,n} \\ ksF_{2,1} & ksF_{2,2} & \dots & ksF_{2,n} \\ \dots & \dots & \dots & \dots \\ ksF_{m,1} & ksF_{m,2} & \dots & ksF_{m,n} \end{bmatrix}, \quad Msf_t = \begin{bmatrix} Msf_{t1} \\ Msf_{t2} \\ \dots \\ Msf_{tm} \end{bmatrix} = \begin{bmatrix} tF_{1,1} & tF_{1,2} & \dots & tF_{1,n} \\ tF_{2,1} & tF_{2,2} & \dots & tF_{2,n} \\ \dots & \dots & \dots & \dots \\ tF_{m,1} & tF_{m,2} & \dots & tF_{m,n} \end{bmatrix}, \quad (2)$$

where Msf_{ksi} and Msf_{ti} stand for the i^{th} KS-based and T-based fluctuation vectors. Then, we infer the two normalized fluctuation matrixes $nMsf_{ks}$ and $nMsf_t$ as,

$$nMsf_{ks} = \begin{bmatrix} nMsf_{ks1} \\ nMsf_{ks2} \\ \dots \\ nMsf_{ksm} \end{bmatrix} = \begin{bmatrix} nksF_{1,1} & nksF_{1,2} & \dots & nksF_{1,n} \\ nksF_{2,1} & nksF_{2,2} & \dots & nksF_{2,n} \\ \dots & \dots & \dots & \dots \\ nksF_{m,1} & nksF_{m,2} & \dots & nksF_{m,n} \end{bmatrix}, nMsf_t = \begin{bmatrix} nMsf_{t1} \\ nMsf_{t2} \\ \dots \\ nMsf_{tm} \end{bmatrix} = \begin{bmatrix} ntF_{1,1} & ntF_{1,2} & \dots & ntF_{1,n} \\ ntF_{2,1} & ntF_{2,2} & \dots & ntF_{2,n} \\ \dots & \dots & \dots & \dots \\ ntF_{m,1} & ntF_{m,2} & \dots & ntF_{m,n} \end{bmatrix}, \quad (3)$$

In which, the normalized elements $nksF_{i,j}$ and $ntF_{i,j}$ are calculated by,

$$nksF_{i,j} = ksF_{i,j} / \max(Msf_{ksi}), \quad ntF_{i,j} = tF_{i,j} / \max(Msf_{ti}), \quad (4)$$

Then, the KST-based fluctuation matrix $fMsf$ can be constructed as,

$$fMsf = \begin{bmatrix} fMsf_1 \\ fMsf_2 \\ \dots \\ fMsf_m \end{bmatrix} = \begin{bmatrix} fF_{1,1} & fF_{1,2} & \dots & fF_{1,n} \\ fF_{2,1} & fF_{2,2} & \dots & fF_{2,n} \\ \dots & \dots & \dots & \dots \\ fF_{m,1} & fF_{m,2} & \dots & fF_{m,n} \end{bmatrix}, \quad (5)$$

In which, each element $fF_{i,j}$ in $fMsf$ is selected from $nMsf_{ks}$ and $nMsf_t$, and then calculated by $fF_{i,j} = \max(nksF_{i,j}, ntF_{i,j})$, $1 \leq i \leq m$, and $1 \leq j \leq (n = \log_2 N)$.

2.2 Construction of the KST-based FCN

To construct a correlation networks for multivariate time series, we first calculate the correlation coefficient $fRsf_{i,j}$ between two vectors $fMsf_i$ and $fMsf_j$ in $fMsf$, and then obtain the fluctuation correlation matrix $fMsfC$ as,

$$fMsfC = \begin{bmatrix} fRsf_1 \\ fRsf_2 \\ \dots \\ fRsf_m \end{bmatrix} = \begin{bmatrix} fRsf_{1,1} & fRsf_{1,2} & \dots & fRsf_{1,m} \\ fRsf_{2,1} & fRsf_{2,2} & \dots & fRsf_{2,m} \\ \dots & \dots & \dots & \dots \\ fRsf_{m,1} & fRsf_{m,2} & \dots & fRsf_{m,m} \end{bmatrix}, \quad (6)$$

In which, each element $fRsf_{i,j}$ in $fMsfC$ can be calculated by,

$$fRsf_{i,j} = \frac{C(fMsf_i, fMsf_j)}{\sqrt{C(fMsf_i, fMsf_i) * C(fMsf_j, fMsf_j)}}, \quad (7)$$

in which, $C(X, Y)$ is the covariance matrix between two vectors X and Y , and formulated as,

$$C(X, Y) = \text{cov}(X, Y) = E[(X - \mu_X) * (Y - \mu_Y)], \quad (8)$$

where X and Y refer to two vectors, $fMsf_i$ and $fMsf_j$ in $fMsf$, respectively, E is the mathematical expectation and $\mu_X = EX$, and $\mu_Y = EY$.

To construct a correlation network that shows the connections between the time series within a multivariate time series, we apply a threshold, $C(\gamma)$, for the correlation coefficients to determine the existence of the connections, and create a

network connection matrix based on the $fMsfc$. γ is a confidence level used in the corresponding T and KS statistics. The connection matrix can be denoted as as,

$$FCN = \begin{bmatrix} FCN_1 \\ FCN_2 \\ \dots \\ FCN_m \end{bmatrix} = \begin{bmatrix} FCN_{1,1} & FCN_{1,2} & \dots & FCN_{1,m} \\ FCN_{2,1} & FCN_{2,2} & \dots & FCN_{2,m} \\ \dots & \dots & \dots & \dots \\ FCN_{m,1} & FCN_{m,2} & \dots & FCN_{m,m} \end{bmatrix}, \quad (9)$$

Where the vector FCN_i refers to the connections (yes/no) between the i^{th} time series and each time series in the multivariate time signals. $FCN_{i,j}$ represents the correlation between each pair of the two time series i and j , which can be denoted as,

$$FCN_{i,j} = \begin{cases} 0, & \text{if } |fRsf_{i,j}| \leq C(\gamma), \\ 1, & \text{if } |fRsf_{i,j}| > C(\gamma), \end{cases} \quad (10)$$

Based on the connection matrix, a connection network can be visualized, with nodes representing the time series and edges added between those nodes whose time series were detected to be correlated.

The construction procedure of the KST-based fluctuation correlation network is described in Algorithm 1.

Algorithm 1. The procedure of FCN construction based on KST method

Require: a set of time series $\{S_1, S_2, \dots, S_m\}$, a threshold of the correlation coefficient $C(\gamma)$

- 1: Divide $\{S_1, S_2, \dots, S_m\}$ into a segment matrix Ms , given the values of segment size N_s and original size N ;
- 2: **for** each time series S_i in $\{S_1, S_2, \dots, S_m\}$ **do** // Build the maximal fluctuation matrix.
- 3:   **for** each segment S_{ij} in S_i **do**
- 4:   Run the KST-based algorithm to detect the maximal fluctuation F_{ij} within S_{ij} ;
- 5:   **end for**
- 6: Assemble the maximal fluctuation vector $fMsfi$ composed of each F_{ij} ;
- 7: **end for**
- 8: Build the maximal fluctuation matrix $fMsf$ composed of each $fMsfi$;
- 9: **for** each element $fMsfi$ in $fMsf$ **do** // Construct the fluctuation correlation network over $\{S_1, S_2, \dots, S_m\}$.
- 10:   **for** each element $fMsfi$ in $fMsf$ **do**
- 11:   Calculate the correlation coefficient $fRsf_{i,j}$ between each pair of $fMsfi$ and $fMsfi$;
- 12:   **if** $|fRsf_{i,j}| > C(\gamma)$
- 13:   Set $FCN_{i,j} = 1$, and add this edge in FCN;
- 14:   **else**
- 15:   Set $FCN_{i,j} = 0$, and add no edge in FCN;
- 16:   **end if**
- 17:   **end for**
- 18: Construct the subnet FCN_i originated from node i ;
- 19: **end for**
- 20: Construct the whole fluctuation correlate network FCN over the set of time series;

3. IMPLEMENTATION AND EXPERIMENTS

To test our method, we created some target correlation networks based on synthesized multivariate time series signals that can reflect the underlying fluctuations in the correlation networks. The constructed correlation networks were compared with the target correlation networks. Similar method was used in [25] to check how the constructed causal

network structures resemble the target causal network. Specifically, there are three steps for a target network construction. First, a multivariable time series $\{S_1, S_2, \dots, S_{Nnod}\}$ was created, and of which each time series was generated by normalized random distribution. Second, each S_i ($i=[1, Nnod]$) was equally divided into N segments. Third, corresponding to the target network with N_{edg} edges, random fluctuations were added into segments in each pair of the correlated nodes.

In our experiments, we compared the topologies of the generated networks from KS, T and KST, by counting the spurious edges and the missing edges against the target correlation network. We then evaluated the constructed network with its false positive (spurious), false negative (missed) and total error rates. The result of 2 sets of our experiments (Exp.1 and Exp.2) are shown in figure 1 and Table 1. In these experiments, we selected the 95% confidence level for both KS and T tests, and an absolute value of 0.9 was chosen for the correlation coefficient threshold. In Exp.1 we synthesized 6 datasets D_1 - D_6 with the fixed segment size 2^7 , which included 6, 9, 12, 15, 18 and 21 nodes respectively. In the Exp2, we used one dataset D_3 with 6 different segment sizes from 2^5 to 2^{10} . The experiment for each dataset with certain segmentation was executed repeatedly 50 times with randomized fluctuation. As a result, a total of 300 correlation networks were created from Exp.1 and Exp.2 respectively.

Figure 1. Results of the synthetic simulations on Exp.1 and Exp.2 by KST, KS and T methods, in terms of the missed and spurious rate, as well as the accuracy. In these histograms, '1', '2', '3' refer to the methods of KST, KS, and T, respectively.

Table 1. Mean analysis for KST, KS and T methods based on Exp.1 and Exp.2.

Items		Missed Rate	Spurious Rate	Accuracy
Methods				
Exp.1	KST	.0132	.0086	.978
	KS	.0180	.0091	.972
	T	.1290	.0137	.857
Exp.2	KST	.0112	.0066	.982
	KS	.0179	.0071	.975
	T	.1179	.0121	.870

Figure 2. The results of FCN construction by KST, KS and T, from the synthetic dataset D_3 and D_6 in Exp.1 with the given values of $N_s=2^5, 2^7$, and 2^9 , respectively.

From the figure and the table, we can see that the result from either of the three models was reasonably good. Not surprised, the method based on KST created the FCNs with smaller false positive and false negative rates than those from KS and T methods. It might be explained by that KST had better ability for detection of the fluctuations from each of the time series segments, which has been addressed in our previous work [19-21]. Some examples of the visualized FCNs are shown in figure 3. We can see the missed and spurious edges from each FCN created based on different method.

4. CONCLUSION

In this paper, a hybrid KST framework is proposed for FCN construction from multivariate time signals. Our method provide a novel strategy for construction of fluctuation correlation networks which may better reflect the associated functional connectivity in the real world applications than the correlation networks created based on the original time series signals. The simulations showed that our KST method constructed FCNs from multivariate time series with smaller error than KS and T did, giving both KS and T also resembled the correlation networks reasonably close to the target networks. Our future work based on this proposal will be testing the KST method for construction of FCNs with real world time series and analyze the associated connections between different objects or functionalities.

Acknowledgment

This paper is supported by National Natural Science Foundation of China (no. 61305081, no. 61104154, and no.61773112) , Specialized Research Fund for Natural Science Foundation of Shanghai (no. 16ZR1401300 and no.16ZR1401200) and Shanghai Postdoctoral Sustentation Fund (no. 12R21410200).

REFERENCES

- [1] S. Dirk, K. Leonard, M. Carsten, B. Dirk, and U. Björn, *Correlation Networks*. John Wiley & Sons, Inc., 305-333, (2007)
- [2] M. Müller-Linow, W. Weckwerth, and M. T. Hütt, "Consistency analysis of metabolic correlation networks," *Bmc Systems Biology*. 1(1), 1-12, (2007)
- [3] T. Aste, T. D. Matteo, M. Tumminello, and R. N. Mantegna, "Correlation filtering in financial time series," *Proceedings of SPIE - The International Society for Optical Engineering*. 5848, 100-109, (2005)
- [4] J. P. Onnela *et al.*, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences of the United States of America*. 104(18), 7332-7336, (2007)
- [5] J. Zhang, . and M. Small, . "Complex network from pseudoperiodic time series: topology versus dynamics," *Physical Review Letters*. 96(12), 238701-238701, (2006)

- [6] G. C. Reinsel, "Elements of Multivariate Time Series Analysis," *Journal of the American Statistical Association*. 92(89), 554-558, (1997)
- [7] B. Schelter *et al.*, "Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction," *Chaos*. 16(1), 298-311, (2006)
- [8] S. Miccichè, F. Lillo, G. Bonanno, and R. N. Mantegna, *Univariate and multivariate statistical aspects of equity volatility*. Springer Japan, (2004)
- [9] R. Albert and R. Albert, "Barab'asi: Statistical mechanics of complex networks," *Rev.mod.phys.* 74(1),2002, (2001)
- [10] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde, "Detecting long-range correlations with detrended fluctuation analysis," *Physica A-statistical Mechanics & Its Applications*. 295, 441-454, (2001)
- [11] I. Barnett and J. P. Onnela, "Change Point Detection in Correlation Networks," *Scientific Reports*. 6(6),443-463, (2016)
- [12] P. Damos, "Using multivariate cross correlations, Granger causality and graphical models to quantify spatiotemporal synchronization and causality between pest populations," *Bmc Ecology*. 16(1),33, (2016)
- [13] E. Siggiridou, D. Kugiumtzis, and V. K. Kimiskidis, "Correlation Networks for Identifying Changes in Brain Connectivity during Epileptiform Discharges and Transcranial Magnetic Stimulation," *Sensors*. 14, 12586-12597, (2014)
- [14] M. Jalili and M. G. Knyazeva, "Constructing Brain Functional Networks from EEG: Partial and Unpartial Correlations," *Journal of Integrative Neuroscience*. 10, (2011)
- [15] C. Feng and B. He, "Construction of complex networks from time series based on the cross correlation interval," *Open Physics*. 15, DOI: <https://doi.org/10.1515/phys-2017-0028>, (2017)
- [16] G. Bonanno, G. Caldarelli, F. Lillo, S. Miccichè, N. Vandewalle, and R. N. Mantegna, "Networks of equities in financial markets," *European Physical Journal B -- Condensed Matter*. 38(2), 363-371, (2004)
- [17] K. Pukenas, "Algorithm for the Detection of Changes of the Correlation Structure in Multivariate Time Series," *Electronics & Electrical Engineering*. 18(8), 53-56, (2012)
- [18] S. Aminikhanghahi and D. J. Cook, "A Survey of Methods for Time Series Change Point Detection," *Knowl Inf Syst*. 51(2), 339-367, (2017)
- [19] J.-P. Qi, Q. Zhang, Y. Zhu, and J. Qi, "A Novel Method for Fast Change-Point Detection on Simulated Time Series and Electrocardiogram Data," *Plos One*. 9(4), e93365, (2014)
- [20] J.-p. Qi, Q. Zhang, F. Pu, and J. Qi, "A Fast Method for Change Point Detection from Large-scale Time Series Based on Haar Wavelet and Binary Search Tree (HWBST)," in *the 33rd Chinese Control Conference*, Nanjing, China, 506-511, (2014)
- [21] J.-p. Qi, J. Qi, F. Pu, and T. Gong, "Multi-Channel Detection for Abrupt Change Based on the Ternary Search Tree and Kolmogorov Statistic Method," in *the 34th Chinese Control Conference*, Hangzhou China, 4968-4973, (2015)
- [22] Y. Yang and H. Yang, "Complex network-based time series analysis," *Physica A Statistical Mechanics & Its Applications*. 387, 1381-1386, (2008)
- [23] Z. Yang, K. T. Fang, and S. Kotz, "On the Student's t-distribution and the t-statistic," *Journal of Multivariate Analysis*. 98(6), 1293-1304, (2007)
- [24] L. M. Surhone, M. T. Tennoe, S. F. Henssonow, and S. S. T-Distribution, "Student's T-statistic," *Betascript Publishing*, (2010).
- [25] S. Acharya and B. S. Lee, "Incremental causal network construction over event streams," *Information Sciences*. 261(5), 32-51, (2014)