Available online at www.sciencedirect.com
ScienceDirect

Transactions of A. Razmadze Mathematical Institute 170 (2016) 352–362

Transactions of A. Razmadze Mathematical Institute
www.elsevier.com/locate/trmi

Original article

A Novel similarity measure based on eigenvalue distribution

Xu Huang[a], Mansi Ghodsi[b], Hossein Hassani[b,*]

[a] *Business School, Bournemouth University, UK*
[b] *Institute for International Energy Studies, Tehran, 1967743 711, Iran*

Received 8 May 2016; received in revised form 2 August 2016; accepted 17 August 2016
Available online 19 September 2016

Abstract

Due to the rapidly increasing interests of effective and efficient data processing, the developments of similarity measure have been significantly expanded. This paper defines the eigenvalue distribution as a criterion of measuring similarity in a multivariate system. The primary evaluations are conducted by simulations with the assistances and comparisons of several empirical statistical tests. Furthermore, the proposed measure is conducted in simultaneous real case scenario by adopting the bootstrap re-sampling technique. It also overcomes the difficulty of different series lengths in the multivariate system. Moreover, it does not have pre-assumptions on distributions, and it can be easily employed and efficiently computed.

Keywords: Similarity measure; Eigenvalue distribution; Singular value decomposition; Multivariate

1. Introduction

The studies of similarity have been overwhelmingly explored and applied in various disciplines on many different formats, for example, numerical values [1,2], images [3,4], genes [5–7], chemical subjects [8–10], words [11,12] and so on. According to [13], the similarity measure is the most essential core element of time series classification and clustering. Therefore, the development of better similarity measure can significantly assist the improvement of data analysis efficiency. According to [14], the similarity measure is closely related to the distance measure, as the distance is defined as a quantitative degree of how far apart two objects are. Consequently, studies of distance and similarity are significantly connected and crucial in terms of solving many pattern recognition related problems, such as clustering technique [15,16], Taxonomy [17,18], image registration [19,20], etc.

As one of the crucial difficulties in similarity measure is that the different types of features are not comparable, this paper proposes the novel similarity measure based on the eigenvalue distribution, which is inspired by the dynamical approach and embedding theorem where a one dimensional time series will be transferred to multidimensional time series in a Hankel matrix. Hankel matrix has many features as a square matrix, where gives a sequence of the

[*] Corresponding author.
E-mail address: hassani.stat@gmail.com (H. Hassani).
Peer review under responsibility of Journal Transactions of A. Razmadze Mathematical Institute.

http://dx.doi.org/10.1016/j.trmi.2016.08.001

one dimensional time series, also defines the dynamical state-space. This paper is the initial attempt of adopting eigenvalue distribution into formulating a similarity measure in the multivariate system. Time series under evaluation are embedded into multidimensional matrices and combined either vertically or horizontally to be transformed into a Hankel matrix, where the eigenvalues can be extracted by Singular Value Decomposition (SVD) technique accordingly. As Aristotle claimed in [21], the Formal Cause is "the account of what-it-is-to-be", or "what makes a thing one thing rather than many things". Based on the "formal cause" claimed by Aristotle, here in this paper, we define the corresponding distribution of extracted eigenvalues as the "formal" criterion for developing a novel similarity measure. The successful implementation of this novel similarity measure can overcome the limitations of nonlinear dynamic, complex fluctuations and the possibility of distinguishing similarity for particular or selected features.

In order to evaluate the reliability of eigenvalue distribution as the similarity measure, three empirical statistical tests together with the real case scenario are overwhelmingly considered. Possible circumstances during the formulation process of the new measure are comprehensively evaluated with brief introductions and comparisons in following sections.

In general, this paper is structured as follows: Section 2 briefly introduce the techniques for obtaining the corresponding eigenvalue distribution. The review of some empirical methods and the formulation of proposed novel similarity measure are listed in Section 3. Section 4 provides the empirical results and evaluations by simulations, whilst the real case scenario results are stated in Section 5. Finally, the discussion and conclusion are summarized in Sections 6 and 7 respectively.

2. Eigenvalue distribution

To overcome the difficulty of existing diverse and incomparable features, the novel similarity measure extracts the corresponding eigenvalue distributions as the formal criterion by considering the elements of time series as a whole without removing any nonlinear or complex features. Note that as the structures of constructing Hankel matrix containing multiple variables differ, including both horizontal and vertical forms.

Consider M time series with different series length N_i $Y_{N_i}^{(i)} = (y_1^{(i)}, \dots, y_{N_i}^{(i)})(i = 1, \dots, M)$. In this case, the standard univariate form can be acquired by setting $M = 1$. Firstly, we transfer a one-dimensional time series $Y_{N_i}^{(i)}$ in to a multidimensional matrix $[X_1^{(i)}, \dots, X_{K_i}^{(i)}]$ with vectors $X_j^{(i)}$ that equals to $(y_j^{(i)}, \dots, y_{j+L_i-1}^{(i)})^T \in \mathbf{R}^{L_i}$, where $L_i (2 \le L_i \le N_i/2)$ is the window length for each series with length N_i and $K_i = N_i - L_i + 1$. We can then get the trajectory matrix $\mathbf{X}^{(i)} = [X_1^{(i)}, \dots, X_{K_i}^{(i)}] = (x_{mn})_{m,n=1}^{L_i,K_i}$ after this step. The above procedure for each series separately provides M different $L_i \times K_i$ trajectory matrices $\mathbf{X}^{(i)} (i = 1, \dots, M)$.

To construct a block Hankel matrix in the vertical form we need to have $K_1 = \cdots = K_M = K$. Accordingly, this version enables us to have various window length L_i and different series length N_i, but similar K_i for all series. The result of this step is the following block Hankel trajectory matrix:

$$\mathbf{X}_V = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(M)} \end{bmatrix}.$$

Note that $\mathbf{X}_V$ indicates that the output of the first step is a block Hankel trajectory matrix formed in a vertical form.

Then, the SVD of $\mathbf{X}_V$ is performed in the following step. Note that the SVD technique is closely related to the Singular Spectrum Analysis technique and its multivariate extension, which have been widely applied in a range of different fields and a multitude of fairly precise results proved it as a powerful and applicable technique [22,29,23–28,30–35]. Denote $\lambda_{V_1}, \dots, \lambda_{V_{L_{sum}}}$ as the eigenvalues of $\mathbf{X}_V \mathbf{X}_V^T$, arranged in decreasing order $(\lambda_{V_1} \ge \cdots \lambda_{V_{L_{sum}}} \ge 0)$ and $U_{V_1}, \dots, U_{V_{L_{sum}}}$, the corresponding eigenvectors, where $L_{sum} = \sum_{i=1}^{M} L_i$. Note also that the structure of the matrix $\mathbf{X}_V \mathbf{X}_V^T$ is as follows:

$$\mathbf{X}_V \mathbf{X}_V^T = \begin{bmatrix} \mathbf{X}^{(1)}\mathbf{X}^{(1)T} & \mathbf{X}^{(1)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(1)}\mathbf{X}^{(M)T} \\ \mathbf{X}^{(2)}\mathbf{X}^{(1)T} & \mathbf{X}^{(2)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(2)}\mathbf{X}^{(M)T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(M)}\mathbf{X}^{(1)T} & \mathbf{X}^{(M)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(M)}\mathbf{X}^{(M)T} \end{bmatrix}.$$

The structure of the matrix $\mathbf{X}_V\mathbf{X}_V^T$ is similar to the variance–covariance matrix in the classical multivariate statistical analysis literature. The matrix $\mathbf{X}^{(i)}\mathbf{X}^{(i)T}$ for the series $Y_{N_j}^{(j)}$, appears along the main diagonal and the products of two Hankel matrices $\mathbf{X}^{(i)}\mathbf{X}^{(j)T}\,(i \neq j)$, which are related to the series $Y_{N_i}^{(i)}$ and $Y_{N_j}^{(i)}$, appears in the off-diagonal. The SVD of $\mathbf{X}_V$ can be written as $\mathbf{X}_V = \mathbf{X}_{V_1} + \cdots + \mathbf{X}_{V_{L_{sum}}}$, where $\mathbf{X}_{V_i} = \sqrt{\lambda_{V_i}}U_{V_i}V_{V_i}^T$ and $V_{V_i} = \mathbf{X}_V^T U_{V_i}/\sqrt{\lambda_{V_i}}$ ($\mathbf{X}_{V_i} = 0$ if $\lambda_{V_i} = 0$).

Moreover, the horizontal form decomposition is proved to produce more reliable and consistent eigenvalue distributions. Note that the eigenvalue distributions by vertically and horizontally formed techniques are both carefully considered and compared (detailed results are available upon request from authors). Hence, all tests in the following sections are based on eigenvalues conducted by decomposition stage of the horizontal form.

3. Similarity measures

The distributions of eigenvalues of the trajectory matrices are here considered as the “formal” criterion of measuring the similarity between two series. The explorations of the significance of the Hankel matrix and its corresponding eigenvalues can be found in many different areas (for example [29,36–39]). In addition, more details about the empirical distribution of the eigenvalues of the Hankel matrix divided by its trace can be found in [40,42,41].

In order to evaluate whether the extracted eigenvalues are similar or not to conclude the similarity between two tested series, three empirical statistical tests (Chi-squared Test, Log-likelihood Goodness of Fit Test and Kolmogorov–Smirnov Test) are adopted. Various distance and similarity measures are comprehensively reviewed and categorized in [14], therefore, we do not reproduce here. Since the proposed similarity measure is expected to have no assumption or limitation on measuring tested series with only the empirical distributions, some tests that are commonly used to evaluate the consistency with the empirical distributions cannot be properly suitable here (i.e. Shapiro–Wilk Test [43], Hellinger Distance [44], Kullback Leibler Divergence [45], Anderson–Darling Test [46]). Therefore, only brief introductions of the suitable empirical statistical tests are provided respectively as follows.

3.1. Similarity measures

In general, coordinates and the cumulative distribution function (CDF) are the most generally accepted concepts to represent the examined subject. We briefly summarize several important and dominant measurements that are referred for formulating the novel similarity measure due to the special feature of eigenvalue distribution.

3.1.1. Chi-squared test

As an improved distance measure comparing to Euclidean distance, the Chi-squared statistic can be simply considered as the summation of squared Euclidean distances of two vectors (by considering them in a n dimensional space domain, where n is the number of observations for both vectors) over the corresponding “coordinates” of the domain vector. The Chi-squared distribution (also known as Helmertian distribution) [47] is one of the most significantly applied probability distributions, and it is most commonly accepted for measuring the distance or similarity level between two probability distributions. Pearson [48] adopted the Chi-squared distribution in the goodness of fit domain and conducted the Chi-squared test, which statistically evaluates the observed data about its goodness of fit level and consistency with an expected distribution. Here in this paper, it is adopted for comparing the eigenvalue distributions of two series (or one examined series with the benchmark population) as evidence of similarity. The Chi-squared statistic formula is:

$$\chi^2(C, E) = \sum_{i=1}^{Z} \frac{(C_i - E_i)^2}{E_i}, \tag{1}$$

where Z is the number of levels of categories; C is the observed frequency and E is the expected count.

Therefore, in terms of Chi-squared test between two tested variables, assume Z_A and Z_B are the number of levels of categorized variables A and B, so the degree of freedom can be calculated by $df = (Z_A - 1) \times (Z_B - 1)$. The expected counts/frequencies is computed by

$$E_{Z_{A,B}} = (C_{Z_A} \times C_{Z_B})/n, \tag{2}$$

where C_Z refers to observed counts at specific level of category and n indicates the total observation number. Consequently, the corresponding Chi-squared statistics is:

$$\chi^2(A, B) = \sum \frac{(C_{Z_{A,B}} - E_{Z_{A,B}})^2}{E_{Z_{A,B}}}. \tag{3}$$

3.1.2. Log-likelihood goodness of fit test

The Log-likelihood Goodness of Fit Test is actually based on the commonly used Chi-squared test statistics in [48]. According to [49], the Log-likelihood statistic formula is:

$$G = 2 \sum_i f_i \cdot ln \left(\frac{f_i}{q_i} \right), \tag{4}$$

where the f_i refers to the observed frequency, whilst q_i indicates the expected frequency. More specifically, the test is adopted for evaluating whether the eigenvalue distribution of the examined series fit well to the eigenvalue distribution of the benchmark series.

3.1.3. Kolmogorov–Smirnov test

The Kolmogorov–Smirnov Test (K–S Test) was firstly proposed in [50]. As a non-parametric statistical test, it quantifies the distance based on the CDF with no assumption about the distribution of data. It can be adopted to examine the similarity level of one distribution to empirical distribution, more importantly, K–S test is also applicable for evaluating the similarity of distributions of two random samples. The K–S test statistic is defined as below, which we mainly follows [51]:

$$D_n = sup_x |F_n(x) - F(x)|, \tag{5}$$

where F refers to the theoretical cumulative distribution function, F_n represents the cumulative distribution up to n observations, sup_x indicates the supremum of the set of distances, and D_n refers to the supremum distance reached up to n observations. In terms of the two-sample case of K–S Test, the corresponding test statistic formula is:

$$D_{n,n'i} = sup_x |F_{1,n}(x) - F_{2,n'}(x)|, \tag{6}$$

note that $F_{1,n}$ and $F_{2,n'}$ are the corresponding distribution function for two tested samples respectively.

Specifically for the proposed similarity measure method based on eigenvalue distribution, two-sample K–S Test is adopted to determine whether the "benchmark" populations created by the dominate series has consistent eigenvalue distribution as the other series.

3.2. Novel similarity measure using eigenvalue distribution

By setting the eigenvalue distribution as our criterion and adopting the empirical methods listed above, the hypotheses of the novel similarity measure are stated as below:

Null hypothesis (H_0): there is no significant difference between the eigenvalue distributions of matrices by two tested series.

Alternative hypothesis (H_a): there is a significant difference between the eigenvalue distributions of matrices by two tested series.

The null hypothesis is rejected when the p-value is less than the 5% significance level, and therefore we conclude that the set of eigenvalues are not similar and consequently two test series are different. While if the p-value is very close to or equal to 1, we conclude that the two tested series are similar as they share very similar or even identical eigenvalue distributions.

As the proposing method of measuring similarity based on eigenvalue distribution is considering a possible implementation of detecting "Formal Cause", different benchmarks of comparison will lead to different results. Consider two random variables X and Y, "how similar is X to Y" and "how similar is Y to X" are two different

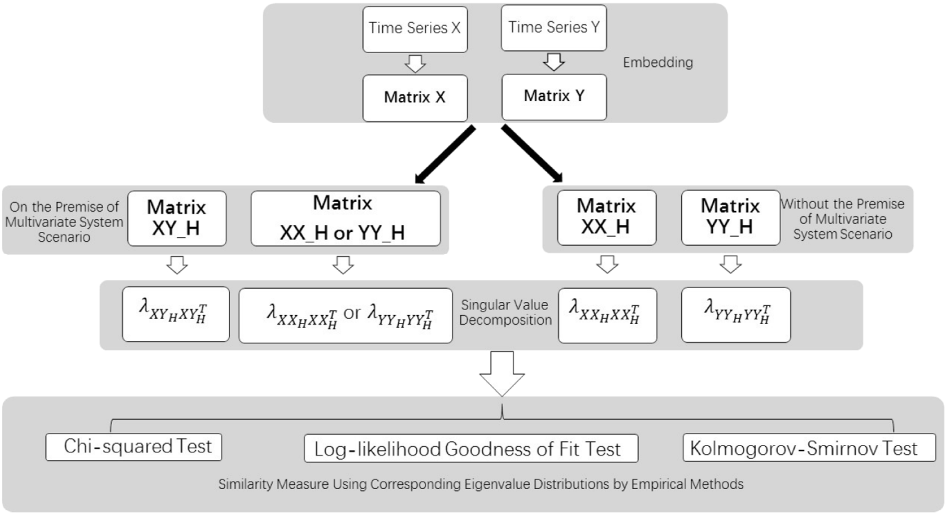
Time Series X
Time Series Y
Embedding
Matrix X
Matrix Y
On the Premise of Multivariate System Scenario
Matrix XY_H
Matrix XX_H or YY_H
Matrix XX_H
Matrix YY_H
Without the Premise of Multivariate System Scenario
$\lambda_{XY_H XY_H^T}$
$\lambda_{XX_H XX_H^T}$ or $\lambda_{YY_H YY_H^T}$
Singular Value Decomposition
$\lambda_{XX_H XX_H^T}$
$\lambda_{YY_H YY_H^T}$
Chi-squared Test
Log-likelihood Goodness of Fit Test
Kolmogorov-Smirnov Test
Similarity Measure Using Corresponding Eigenvalue Distributions by Empirical Methods

Fig. 1. The flowchart of the novel similarity measure using eigenvalue distribution.

questions, especially when the distribution of eigenvalues is the criterion. We should not expect exactly "same" results when we compare X to Y and Y to X, while the expected final outcomes that define "similar" or "different" should not vary. For instance, if the principle is to answer the question of how similar is Y to X, the eigenvalue distribution by corresponding matrix ($\mathbf{XY}_H$ or $\mathbf{XX}_H$ determined by with or without the premise of multivariate system) will be considered as the "benchmark" for further evaluation. Hence, if the other eigenvalue distribution by $\mathbf{XX}_H$ or $\mathbf{YY}_H$ (determined by with or without the premise of multivariate system respectively) is statistically similar with the "benchmark" eigenvalue distribution, Y will then concluded as similar with X.

Moreover, in order to ensure the consistency and comparability, the default window length is set as about 1/10 of the time series length. This will be fairly number to include almost all significant eigenvalues without containing too much unimportant ones. With a relatively larger window length, the information will be split either flatly or partly flatly by more eigenvalues, and the differences will be split to be less significant to be identified; in contrast, a smaller window length will result in the fewer amount of eigenvalues with more significant differences for all or some of the eigenvalues. Without considering the consistency to be comparable, the most proper window length will be selected heavily depends on the feature of the series being analyzed with the principle of relatively maximizing the significant information with possibly small number of eigenvalues.

A flowchart is provided in Fig. 1 that briefly summarizes the formulation and evaluation process of this proposing similarity measure. Note that in terms of simulation, corresponding process is repeated 1000 times respectively, and the population of tested series are generated by involving random white noises that being maintained at about 10% of the range of tested series.

The similarity measure is firstly built on the premise of multivariate system with the benchmark series as the dominant role. Therefore, we evaluate the similarity of multivariate system formed by X and Y by comparing it to the benchmark multivariate system formed by X and X or Y and Y respectively (determined by which series is considered as the benchmark series). This will be considered as the scenario of on the premise of multivariate system.

Another question raise here is that we can only compare the system of X and X with Y and Y. This refers to the scenario of without the premise of multivariate system. Note that all evaluations will be performed on the corresponding eigenvalue distributions generated by the systems formed respectively. The detailed test results of simulations with and without the premise scenarios will be separately presented in the following sections.

4. Empirical results

Three statistical tests are adopted for evaluating this novel similarity measure and examining the similarity measure criterion of eigenvalue distribution, which are briefly introduced previously: Chi-squared Test, Log-likelihood Goodness of Fit Test, Kolmogorov–Smirnov Test. In order to evaluate the performance of the proposed method, various types of simulated series are tested by being separated into two groups of circumstances: the similar group and the different group, additionally the different choices of "benchmark" are also considered in each group. The test results are summarized in Tables 1 and 2 by each empirical statistical method. Thus, the robustness of accepting eigenvalue distribution as similarity measure criterion are preliminarily examined, followed by the tests under simultaneous real case scenario by employing bootstrap re-sampling technique. We have managed to obtain consistently promising results as simulative expectations, which convincingly prove the consistent, robust performances of this novel similarity measure on several different types of simulated series. The initials of various types of generated series are listed below for the sake of simplifying the expressions:

1.	WN	White noise.
2.	UD[0, 1]	Uniform distribution series [0, 1].
3.	UD[−1, 1]	Uniform distribution series [−1, 1].
4.	EP[1]	Exponential distribution series rate 1.
5.	SINE[−1, 1]	Sine wave series [−1, 1].

4.1. On the premise of multivariate system scenario

Regarding the scenario of on the premise of multivariate system, we evaluate the similarity of eigenvalue distributions extracted from the matrices $\mathbf{XY}_H$ and $\mathbf{XX}_H$ (or $\mathbf{YY}_H$ determined by which series is considered as the benchmark series), respectively. Note that $\mathbf{XY}_H$ is created from two time series X_N and Y_N simultaneously, and $\mathbf{XX}_H$ (or $\mathbf{YY}_H$) is formed by X_N (or Y_N) with itself respectively. The corresponding test results of eigenvalue distributions as novel similarity measure by three different empirical methods are summarized in Table 1. Note that the bold number indicates the best performance option in corresponding comparable level.

The Chi-squared test results show positive outcomes as expected for the "similar" group on both numbers of observations scenarios, whilst in terms of the "different" group, the tests can perform better for longer series. However, there are still significantly unexpected results (p-value is close to 1) for the UD[0,1] & EP[1] and UD[−1,1] & SINE[−1,1] combinations, especially the results vary greatly for the UD[0,1] & SINE[−1,1] and EP[1] & SINE[−1,1] cases. As mentioned earlier, the population for comparison is created by the "benchmark" series, therefore differences are expected when switching the "benchmark" series, however, opposite results for the same pair of series are not robust as expected, and it is even worse than the cases of indicating "similar" for the groups that are expected to be "different".

In terms of the log-likelihood goodness of fit test results, expected results for the "similar" group are confirmed in accordance with the simulation results. P-values are equal to 1, which indicate that it is almost 100% sure to accept the null hypothesis, therefor very similar or identical eigenvalue distributions prove the expected conclusion of "similar". Regarding the expected to be "different" group, both long and short series length, 1000 and 100 observations, show generally consistent significant results, except the UD[0,1] & EP[1] combination. Since UD[0,1] and EP[1] indeed show similar eigenvalue distributions and the differences are between the tails, the log-likelihood goodness of fit test is not sensitive for detecting differences of distributions with flat tails. However, the advantage of this test can be noticed in the shorter length of observation scenario; the results are almost stable and consistent with the expected results of highly significant statistics.

Test results of K–S test show positive results as expected for the "similar" group on both numbers of observations scenarios. In terms of $N = 100$ case for "different" group of combinations, only the UD[0,1] & UD[−1,1] combination can be detected with 10% of significance level, however, the differences between switching dominant series to create "benchmark" populations are not significant. Comparing to the results of previous tests, the inconsistency is worse than less sensitivity of accurate detection, it has to be noticed that the two sample K–S test shows great performance on consistency and stability, even in the quite unstable and greatly varied scenarios that other tests cannot even provide uniformed results. In addition, for the "different" group with $N = 1000$ case, almost all results are as expected to be significant (majority is under 5%, only a few are under 10%). Note that the EP[1] & SINE[−1,1] combination is the only one that K–S test could not detect significantly, and this is mostly because that K–S test is not that much sensitive to the differences at tail, also the natural character of eigenvalue distribution

Table 1
Similarity measure evaluation by three different tests on simulated groups of series on the premise of multivariate system scenario.

			Chi-squared test				Log-likelihood GOF test				K–S test			
			$N = 100$		$N = 1000$		$N = 100$		$N = 1000$		$N = 100$		$N = 1000$	
			$L = 10$		$L = 100$		$L = 10$		$L = 100$		$L = 10$		$L = 100$	
			p-value		p-value		p-value		p-value		p-value		p-value	
	X	Y	Y→X	X→Y	Y→X	X→Y	Y→X	X→Y	Y→X	X →Y	Y→X	X→Y	Y→X	X→Y
Similar	UD[0,1]	UD[0,1]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	UD[−1,1]	UD[−1,1]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	EP[1]	EP[1]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	SINE[−1,1]	SINE[−1,1]	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Different	UD[0,1]	UD[−1,1]	0.14	**0.00**	0.00	0.00	**0.04**	0.04	0.00	0.00	0.07	0.05	**0.00**	**0.00**
	UD[0,1]	EP[1]	0.76	0.98	0.88	1.00	0.81	1.00	0.99	1.00	**0.77**	**0.65**	**0.01**	**0.01**
	UD[0,1]	SINE[−1,1]	0.98	0.35	1.00	0.00	**0.00**	**0.00**	**0.00**	**0.00**	0.76	0.89	0.02	0.01
	UD[−1,1]	EP[1]	**0.01**	0.88	0.00	0.01	0.05	**0.35**	0.00	0.00	0.49	0.65	**0.00**	**0.00**
	UD[−1,1]	SINE[−1,1]	1.00	1.00	1.00	0.99	**0.00**	**0.00**	**0.00**	**0.00**	0.11	0.41	0.10	0.10
	EP[1]	SINE[−1,1]	1.00	0.20	1.00	0.00	**0.00**	**0.00**	**0.00**	**0.00**	0.45	0.60	0.56	0.53

Table 2
Similarity measure evaluation by three different tests on simulated groups of series without the premise of multivariate system scenario.

			Chi-squared test		Log-likelihood GOF test		K–S test	
			$N = 100$	$N = 1000$	$N = 100$	$N = 1000$	$N = 100$	$N = 1000$
			$L = 10$	$L = 100$	$L = 10$	$L = 100$	$L = 10$	$L = 100$
	X	Y	p-value	p-value	p-value	p-value	p-value	p-value
Similar	UD[0,1]	UD[0,1]	0.99	1.00	0.99	1.00	0.99	0.99
	UD[−1,1]	UD[−1,1]	0.99	1.00	0.99	1.00	0.99	0.98
	EP[1]	EP[1]	0.99	1.00	0.99	1.00	0.98	0.93
	SINE[−1,1]	SINE[−1,1]	0.99	1.00	0.99	1.00	0.99	0.99
Different	UD[0,1]	UD[−1,1]	0.01	0.00	**0.01**	0.00	0.03	**0.00**
	UD[0,1]	EP[1]	0.72	0.73	0.72	0.64	**0.61**	**0.00**
	UD[0,1]	SINE[−1,1]	**0.52**	0.45	0.54	0.49	0.76	**0.01**
	UD[−1,1]	EP[1]	0.22	0.00	**0.18**	0.00	0.23	**0.00**
	UD[−1,1]	SINE[−1,1]	**0.96**	0.46	0.96	0.50	0.98	**0.03**
	EP[1]	SINE[−1,1]	**0.58**	**0.49**	0.59	0.50	0.99	0.57

for both types of series vary at the tail part with increasing differences when the window length of structuring matrix increases.

4.2. *Without the premise of multivariate system scenario*

In terms of the scenario without the premise of multivariate system, the similarity measure is performed on the eigenvalue distributions extracted from the matrices $\mathbf{XX_H}$ and $\mathbf{YY_H}$ respectively. To be consistent with the previous evaluation process, we consider both similar and different groups of series and evaluate the performance of similarity measure by 1000 time simulations. Note that this time there is no premise of a multivariate system, therefore, the evaluation by simulated series will have no assumption on benchmark series. Hence, for each pair of series, there is only one test statistic conducted. The default number of observation is 1000 and default window length is 100. All statistical tests results are listed in Table 2. Note that the bold number indicates the best performance option in corresponding comparable level.

It is worth to be noted that due to the algorithm of applying Chi-square test and Log-likelihood goodness of fit test for two sample test, it is necessary to define one of the tested series as dominant series and re-scale the assumption of distribution in the first place for the further tests. Consequently, for the scenario of without the premise of multivariate system, the simulations of 1000 times are equally shared by both series in one pair of tested series. Therefore, both series have same quantity of chances to be the dominant series to re-scale the assumption distribution. K–S test do not have assumptions on any distribution, hence simulations for two sample test of K–S test here do not have significant difference comparing to the corresponding process of previous scenario on the premise of multivariate system.

According to Table 2, all statistical tests provide consistent results on both short and long series for the similar group, which all show p-value nearly equal or identical to 1. Hence it indicates significantly similar eigenvalue distributions consequently the similarity between tested series. However, in terms of the different group, both Chi-squared test and Log-likelihood goodness of fit test could not detect most of the differences properly except the UD[0,1] & UD[−1,1] and UD[−1,1] & EP[1] combinations. It is mostly because of the variation and instability caused by switching dominant series for re-scale distribution assumption. Even for the longer series case, most of the results get smaller p-values (which indicates different eigenvalue distributions), they are still not significant enough as we expected for the generated different group. K–S test is proved to outperform the other two tests for the long series case, also it can accurately detect the similarity or differences for both simulated groups. Even for the short series case, the results of K–S test are fairly close to the results of the other tests. Unlike the previous test results of log-likelihood goodness of fit test, it does not show good performance on short series this time. In general, by considering the scenario without the premise of a multivariate system, the K–S test is confirmed again as the most proper statistics to be adopted for the new similarity measure based on eigenvalue distribution.

5. Similarity measure in simultaneous real case scenario by bootstrap re-sampling

Based on previous evaluations of eigenvalue distribution as similarity measure criterion by simulations, it can be summarized that the eigenvalue distribution can be considered as a proper criterion of measure similarity by adopting proper statistical test; K–S test outperforms others in the large data size domain with consistent results as simultaneously expected.

Considering the real case scenario, data can be assumed to be formed by signal and noise. Therefore, we cannot simulate noises to form and produce the population of dominate series as the benchmark to measure similarity. Consequently, we adopt bootstrap re-sampling technique [52] to conduct the population of dominate series with specific confidence level and evaluate how similar the other tested variable is to the benchmark population under the specific confidence level circumstance. Note that the newly proposed method can certainly be performed without any re-sampling process if there are already clear information of its population. The corresponding population will only be generated by re-sampling for obtaining the information of its population. Due to the nature of similarity we mentioned previously, the similarity level of X to Y and Y to X are two different questions regarding the differences of the benchmark. Consequently, the re-sampling process will consider two different cases by choosing different original series to create the population.

A flowchart is provided in Fig. 2 that briefly summarizes the formulation process under the simultaneous real case scenario by bootstrap re-sampling. For instance, when the principle is to obtain the population of benchmark series X, thus, the population of $\lambda_{XY_H XY_H^T}$ or $\lambda XX_H XX_H^T$ (determined by with or without the premise of multivariate system) are conducted, which are formed by eigenvalues distributions within specific confidence interval of K–S statistics. Therefore, if the confidence level is fixed as 95%, we can conduct the population of eigenvalue distributions that indicate significantly 95% similarity level with benchmark series. To this end, we can evaluate the other series by comparing its corresponding K–S statistics with the range of K–S statistics by the population. Hence, we can identify the similarity level respectively with necessary adjustment of confidence level in the bootstrap re-sampling stage. The results by representative simultaneous groups of series are provided in Table 3. In terms of the similar group, the similar group shows consistent results for both short and long series, in which, 95% significant level indicates tested series share at least 95% of similarity based on the eigenvalue distributions from the corresponding matrices. According to the previous evaluations of K–S test on short and long series for different group, we here only consider to evaluate the performance on long series in accordance to its previous promising results in simulations (symbol \ for short series in Table 3). The 5% significant level refers to that the test statistics does not fit even when the confidence level of bootstrap re-sampling is set as 5%. This significantly indicates that tested series can be considered different as they are not similar even for 5% significant level.

6. Discussion

Although as a novel similarity measure based on eigenvalue distribution with proven robustness and consistent performances, it is also certain that it is still the beginning of developing this new measure. The types of series in simulations are relatively limited, and there are still numerous choices of more complex series or combinations of

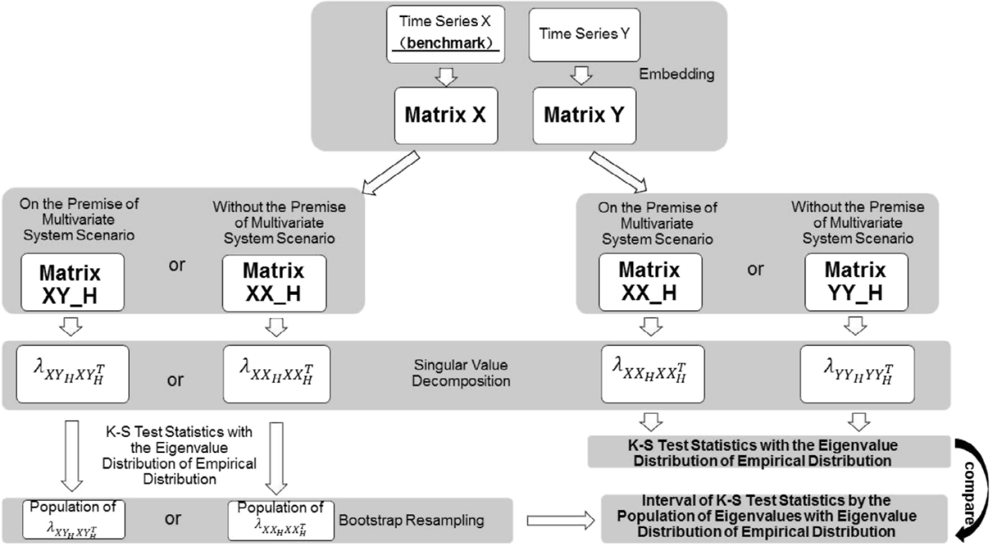
Time Series X (benchmark)
Time Series Y
Embedding
Matrix X
Matrix Y
On the Premise of Multivariate System Scenario
Without the Premise of Multivariate System Scenario
Matrix XY_H
or
Matrix XX_H
On the Premise of Multivariate System Scenario
Without the Premise of Multivariate System Scenario
Matrix XX_H
or
Matrix YY_H
$\lambda_{XY_H XY_H^T}$
or
$\lambda_{XX_H XX_H^T}$
Singular Value Decomposition
$\lambda_{XX_H XX_H^T}$
$\lambda_{YY_H YY_H^T}$
K-S Test Statistics with the Eigenvalue Distribution of Empirical Distribution
K-S Test Statistics with the Eigenvalue Distribution of Empirical Distribution
compare
Population of $\lambda_{XY_H XY_H^T}$
or
Population of $\lambda_{XX_H XX_H^T}$
Bootstrap Resampling
Interval of K-S Test Statistics by the Population of Eigenvalues with Eigenvalue Distribution of Empirical Distribution

Fig. 2. The flowchart of the simultaneous real case scenario by bootstrap re-sampling.

Table 3
Simultaneous real case similarity measure results by bootstrap re-sampling.

			$N = 100$ $L = 10$				$N = 1000$ $L = 100$			
			Y to X		X to Y		Y to X		X to Y	
	X	Y	Y/N	Sig level	Y/N	Sig level	Y/N	Sig level	Y/N	Sig level
Similar	UD[0,1]	UD[0,1]	✓	95%	✓	95%	✓	95%	✓	95%
	UD[−1,1]	UD[−1,1]	✓	95%	✓	95%	✓	95%	✓	95%
	EP[1]	EP[1]	✓	95%	✓	95%	✓	95%	✓	95%
Different	UD[0,1]	UD[−1,1]	\	\	\	\	✓	5%	✓	5%
	UD[0,1]	EP[1]	\	\	\	\	✓	5%	✓	5%
	UD[−1,1]	EP[1]	\	\	\	\	✓	5%	✓	5%

Note: ✓indicates the result is correctly proved by the measure.

series haven not been explored. The bootstrap re-sampling by K–S statistics for some real data (especially large size of data that is much longer than the default 1000 observations in simulation) may take a longer time of calculation, which makes it crucial to find a more straight forward process to identify the population information as the benchmark. Also, the performance in short series is not as good as its effort on long series. However, there are also numerous possibilities to improve this novel measure further: more representative data patterns, more types of noises with different levels of variations and more options of window lengths are planned to be explored as the second stage of improving this new measure; in terms of time series with different frequencies, it can also provide possible solution by adopting SSA technique with specific modification accordingly; one significant implementation area of similarity measure is time series classification, therefore, the evaluations of its performances on classifications of some empirical data are in process.

7. Conclusion

In general, we overcome the difficulties and develop a novel similarity measure based on eigenvalue distribution by combining the SVD technique. It is the initial attempt of adopting this technique in terms of the similarity measure. The evaluation results are promising and robust as we have considered many possible circumstances in the formulation process. We have examined the robustness of adopting eigenvalue distribution as proper criterion of measuring similarity; additionally, we have found that K–S test outperforms others in the large data size domain with consistent results as simultaneously expected. Furthermore, the simultaneous real case scenario is evaluated by adopting the bootstrap re-sampling technique to prevent the possible impacts during the process of creating benchmark population. Consistent results are achieved in the simultaneous real case scenario indicating the robust performance of distinguishing various “similar” or “different” groups of series.

This novel similarity measure can work properly on long series, and it does not require any assumption of distributions during the measuring process. The computation is reasonably efficient and can be easily employed by modifying currently available R packages. By considering eigenvalue distribution as the criterion of similarity measure, the amount of computation is significantly reduced for large data set. More importantly, this novel similarity measure can work with time series with different lengths and still identify the significant features for evaluations. Furthermore, the signal and noise of time series are considered as a whole without one fixed model. In brief, this novel similarity measure contributes to providing a measurement that has no limitations of series length, series with nonlinear features or complex fluctuations, series sharing both signal and noises as similarities, etc. It is absolutely worth looking forward to its developments and performance of implementations on various disciplines in the close future.

References

[1] H.B. Mitchell, On the dengfeng–chuntian similarity measure and its application to pattern recognition, Pattern Recognit. Lett. 24 (16) (2003) 3101–3104.

[2] W.L. Hung, M.S. Yang, Similarity measures of intuitionistic fuzzy sets based on hausdorff distance, Pattern Recognit. Lett. 25 (14) (2004) 1603–1611.

[3] A. Roche, G. Malandain, X. Pennec, N. Ayache, The correlation ratio as a new similarity measure for multimodal image registration, in: Medical Image Computing and Computer-Assisted Interventation, MICCAI’98, Springer, Berlin Heidelberg, 1998, pp. 1115–1124.

[4] C. Yang, R. Duraiswami, L. Davis, Efficient mean-shift tracking via a new similarity measure, in: Computer Vision and Pattern Recognition, IEEE Computer Society Conference, vol. 1, 2005, pp. 176–183.

[5] P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble, Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation, Bioinformatics 19 (10) (2003) 1275–1283.

[6] R. Balasubramaniyan, E. Hullermeier, N. Weskamp, J. Kamper, Clustering of gene expression data using a local shape-based similarity measure, Bioinformatics 21 (7) (2005) 1069–1077.

[7] C.O. Daub, R. Steuer, J. Selbig, S. Kloska, Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data, BMC Bioinformatics 5 (1) (2004) 1.

[8] J.M. Barnard, G.M. Downs, Clustering of chemical structures on the basis of two-dimensional similarity measures, J. Chem. Inf. Comput. Sci. 32 (6) (1992) 644–649.

[9] N. Nikolova, J. Jaworska, Approaches to measure chemical similarity–a review, QSAR Comb. Sci. 22 (9–10) (2003) 1006–1026.

[10] R. Carbo, L. Leyda, M. Arnau, How similar is a molecule to another? An electron density measure of similarity between two molecular structures, Int. J. Quantum Chem. 17 (6) (1980) 1185–1189.

[11] M. Sahami, T.D. Heilman, A web-based kernel function for measuring the similarity of short text snippets, in: Proceedings of the 15th international conference on World Wide Web, AcM, 2006, pp. 377–386.

[12] A. Huang, Similarity measures for text document clustering, in: Proceedings of the Sixth New Zealand Computer Science Research Student Conference, NZCSRSC2008, Christchurch, New Zealand, 2008, pp. 49–56.

[13] J. Serrà, J.L. Arcos, An empirical evaluation of similarity measures for time series classification, Knowl.-Based Syst. 67 (2014) 305–314.

[14] S.-H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, City 1 (2) (2007) 1.

[15] R.A. Jarvis, E.A. Patrick, Clustering using a similarity measure based on shared near neighbors, IEEE Trans. Comput. 100 (11) (1973) 1025–1034.

[16] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.

[17] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007, 1995.

[18] D.K. Lin, An information-theoretic definition of similarity, in: ICML, Vol. 98, 1998, pp. 296–304.

[19] A. Roche, G. Malandain, X. Pennec, N. Ayache, The correlation ratio as a new similarity measure for multimodal image registration, in: Medical Image Computing and Computer-Assisted Interventation, MICCAI’98, Springer, 1998, pp. 1115–1124.

[20] G.P. Penney, J. Weese, J.A. Little, P. Desmedt, D.L. Hill, D.J. Hawkes, A comparison of similarity measures for use in 2-d-3-d medical image registration, IEEE Trans. Med. Imaging 17 (4) (1998) 586–595.

[21] Aristotle. (300s B.C.) Physics, 2.3, 194b17–195a4.

[22] K. Patterson, H. Hassani, S. Heravi, A. Zhigljavsky, Forecasting the final vintage of the industrial production series, J. Appl. Stat. 38 (10) (2010) 2183–2211.

[23] Y. Mohammad, T. Nishida, Discovering causal change relationships between processes in complex systems, in: System Integration (SII), 2011 IEEE/SICE International Symposium on, IEEE, 2011, pp. 12–17.

[24] X.J. Zhao, P.J. Shang, Q.Y. Jin, Multifractal detrended cross-correlation analysis of chinese stock markets based on time delay, Fractals 19 (3) (2011) 329–338.

[25] M. Kapl, W.G. Müller, Prediction of steel prices: a comparison between a conventional regression model and mssa, Stat. Interface 3 (2010) 369–275.

[26] V. Oropeza, M. Sacchi, Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis, Geophysics 76 (3) (2011) V25–V32.

[27] A. Groth, M. Ghil, Multivariate singular spectrum analysis and the road to phase synchronization, Phys. Rev. E 84 (3) (2011) 036206.

[28] H. Hassani, S. Heravi, A. Zhigljavsky, Forecasting uk industrial production with multivariate singular spectrum analysis, J. Forecast. 32 (5) (2013) 395–408.

[29] H. Hassani, A.S. Soofi, A. Zhigljavsky, Predicting inflation dynamics with singular spectrum analysis, J. Roy. Statist. Soc. Ser. A 176 (3) (2013) 743–760.

[30] H. Hassani, X. Huang, R. Gupta, M. Ghodsi, Does sunspot numbers cause global temperatures? A reconsideration using non-parametric causality tests, Physica A 460 (2016) 54–65. http://dx.doi.org/10.1016/j.physa.2016.04.013.

[31] H. Hassani, Singular spectrum analysis: methodology and comparison, J. Data Sci. 5 (2) (2007) 239–257.

[32] D.S. Broomhead, G.P. King, Extracting qualitative dynamics from experimental data, Physica D 20 (2) (1986) 217–236.

[33] N. Golyandina, V. Nekrutkin, A. Zhigljavsky, Analysis of time series structure: SSA and related techniques, CRC Press, 2010.

[34] D. Danilov, A. Zhigljavsky, Principal Components of Time Series: the ‘Caterpillar’ Method, University of St. Petersburg, St. Petersburg, 1997, pp. 1–307.

[35] H. Hassani, R. Mahmoudvand, Multivariate singular spectrum analysis: A general view and new vector forecasting approach, Int. J. Energy Statist. 1 (1) (2013) 55–83.

[36] H. Hassani, S. Heravi, A. Zhigljavsky, Forecasting european industrial production with singular spectrum analysis, Int. J. Forecast. 25 (1) (2009) 103–118.

[37] S. Sanei, T. K.-M. Lee, V. Abolghasemi, A new adaptive line enhancer based on singular spectrum analysis, IEEE Trans. Biomed. Eng. 59 (2) (2012) 428–434.

[38] S. Sanei, M. Ghodsi, H. Hassani, An adaptive singular spectrum analysis approach to murmur detection from heart sounds, Med. Eng. Phys. 33 (3) (2011) 362–367.

[39] H. Hassani, D. Thomakos, A review on singular spectrum analysis for economic and financial time series, Stat. Interface 3 (3) (2010) 377–397.

[40] H. Hassani, N. Alharbi, M. Ghodsi, A study on the empirical distribution of the scaled hankel matrix eigenvalues, J. Adv. Res. (2014).

[41] M. Ghodsi, N. Alharbi, H. Hassani, The empirical distribution of the singular values of a random hankel matrix, Fluct. Noise Lett. 14 (3) (2015) 1550027.

[42] H. Hassani, N. Alharbi, M. Ghodsi, A short note on the pattern of the singular values of a scaled random hankel matrix, Int. J. Appl. Math. 27 (3) (2014) 237–243.

[43] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika (1965) 591–611.

[44] A. Bhattacharyya, On a measure of divergence between two multinomial populations, Sankhyā (1946) 401–406.

[45] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. (1951) 79–86.

[46] T.W. Anderson, D.A. Darling, Asymptotic theory of certain ”goodness of fit” criteria based on stochastic processes, Ann. Math. Stat. (1952) 193–212.

[47] F.R. Helmert, Über die wahrscheinlichkeit der potenzsummen der beobachtungsfehler, Z. Math. u. Phys. 21 (1876) 192–218.

[48] K. Pearson, X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, The London, Edinburgh, and Dublin Phil. Mag. and J. Sci. 50 (302) (1900) 157–175.

[49] R.R. Sokal, F.J. Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research, second ed., 1981.

[50] A.N. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione, in: Giorn. Ist. Ital. Attuari, vol. 4, 1933.

[51] H. Hassani, E.S. Silva, A kolmogorov–smirnov based test for comparing the predictive accuracy of two sets of forecasts, Econometrics 3 (3) (2015) 590–609.

[52] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Statist. 7 (1) (1979) 1–26.