ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/312262279

# A novel mutual association measure based on eigenvalue-based criterion

**Article** · December 2016

DOI: 10.1142/S2335680416500174

CITATIONS

0

READS

2

**2 authors**, including:

Xu Huang

De Montfort University

**8** PUBLICATIONS **6** CITATIONS

SEE PROFILE

All content following this page was uploaded by Xu Huang on 15 January 2017.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

International Journal of Energy and Statistics
Vol. 4, No. 4 (2016) 1650017 (18 pages)
© Institute for International Energy Studies
DOI: 10.1142/S2335680416500174

# A novel mutual association measure based on eigenvalue-based criterion

X. Huang*,‡ and M. Ghodsi†

*Business School, HeNan University of Economics and Law, 450000, China

†Department of Statistics, PNU University, Tehran, 19395-4697, Iran

‡huangxu1118@gmail.com

Received 2 September 2016
Revised 27 October 2016
Accepted 2 November 2016
Published 31 December 2016

The study of association has been one of the domain subjects for the research of multivariate system, which is also known as relationship detection, correlation analysis or dependency identification for a number of different disciplines like economics, biology, chemistry, image or signal processing, etc. The significant effects of association study indicate that the development and improvement of association measure can directly influence the accuracy and possibility of identifying possible relationships among random variables in multivariate system. This paper aims at seeking novel association measure by combining eigenvalue-based criterion. We conduct a brief review of the well established association measures to date, also the formulation process of the novel method is proposed. Furthermore, the novel association measure is compared with the empirical methods by both simulations and a case of real data. The robust performance of this novel method is proved with consistently significant outcomes achieved on linear relationship as well as the nonlinear patterns, which further indicates valuable potentials on nonlinear relationship detection and association measure on complex systems.

*Keywords*: Mutual association; eigenvalue; eigenvalue-based distance; singular spectrum analysis.

## Nomenclature

*BRT* : Europe Brent Spot Price.
*DisCorr* : Distance Correlation.
*HoefD* : Hoeffding's D Test.
*MI* : Mutual Information.
*MIC* : Maximal Information Coefficient.

*MSSA* : Multivariate Singular Spectrum Analysis.
*SSA* : Singular Spectrum Analysis.
*SVD* : Singular Value Decomposition.

## 1. Introduction

Association can be briefly explained as representation of any relationships, or measurement of independency between tested subjects. The studies of association in statistical aspect can be tracked back to over one century ago. As one of the domain subjects in the study of multivariate system, the study of association, or identically named correlation analysis and dependency identification have been developed and applied across subjects on various disciplines, for example, economics [1, 2], social science [3], chemistry [4], biology [5], etc. To date, there are several established association measures with advantages on either linear or nonlinear relationship detection, for instances, Pearson [6], Spearman [7], Kendall [8, 9], Hoeffding's D [10], Distance Correlation [11], Mutual Information [13] and Maximal Information Coefficient [14]. However, there are still numerous possibilities for further improvements as none of these measures can master significant performances for the detection of all possible relationships in a broad sense.

In this paper, we are seeking to develop a novel association measure that is more sensitive on detecting nonlinear or complex associations without losing the ability on basic linear association detection. The technique adopted is the powerful advanced time series analysis technique — Singular Spectrum Analysis (SSA), which has been applied and proved with promising performances on time series analysis, forecasting, denoising and multivariate analysis across various disciplines [15–19]. As an nonparametric time series analysis technique, SSA has the advantage of assumption free and great sensitivity on nonlinear fluctuation and complex pattern detection. This paper is the first attempt of developing SSA technique on association study from a multivariate system aspect. We adopt the eigenvalue-based distance [16] and propose the concept of mutual association measure by considering eigenvalue-based distance as the criterion for measurement.

In order to evaluate the reliability of this novel association measure, a few well established association measures are summarized and overwhelmingly considered as comparison. The performances of both empirical and novelly proposed association measures are evaluated by comprehensive simulations involving representative linear and nonlinear relationships. Furthermore, one real data case is conducted to evident on the robust performance of the novel mutual association measure in actual application scenario.

In genera, this paper is formed as follows: Sec. 2 provides a brief review of several well established association measures expertising in linear or nonlinear relationship detection respectively; the development and formulation process of the novel mutual association measure together with the brief introductions of adopted technique are summarized in Sec. 3; Sec. 4 concludes the evaluations of both empirical and novel

association measures by simulations; one case of real data application is furthermore conducted and evaluated in Sec. 5; finally the conclusion is summarized in Sec. 6.

## 2. Brief Review of Empirical Association Measures

Here in this section, we briefly summarize a few empirical association measures that are generally accepted and well established in literatures by classifying them into linear and nonlinear domains as follows.

### 2.1. *Linear correlation*

#### 2.1.1. *Pearson correlation coefficient*

The Pearson correlation coefficient [6] has been generally accepted as the most significant and well known measurement index to examine the linear dependency level or correlation relationship between tested variables. The calculation process is easy and it has been applied for the majority of practical implementations in terms of association study. The Pearson correlation coefficient, $\rho$, between two random variables $X$ and $Y$ each containing $n$ observations is defined as:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{1}$$

where $E$ is the expected value operator, $\mu_X$, $\sigma_X$ and $\mu_X$, $\sigma_Y$ are expected value and standard deviation of random variables $X$ and $Y$, respectively.

Consequently, under null hypothesis circumstance, the Pearson correlation coefficient $\rho$ computed by the formula above follows the $t$-distribution with degree of freedom of $n-2$. The $t$ statistic is calculated by:

$$t = \frac{\rho(X,Y)\sqrt{n-2}}{\sqrt{1-\rho^2(X,Y)}}. \tag{2}$$

Pearson correlation coefficient $\rho$ satisfies $-1 \leq \rho \leq 1$ with the special cases of perfect linear dependence that equals to $-1$ or 1. It measures the direction and the dependence level between two tested variables in a linear domain. However, if two variables cannot be identified correlation by Pearson correlation, we cannot deny the possibility that they are associated in a nonlinear domain.

#### 2.1.2. *Spearman rank correlation*

Spearman rank correlation [7] is another well accepted measure of dependency level between two variables. It is a nonparametric test and used ranked values to evaluate the association level based on the underlying assumption of a monotonic relationship. The monotonic relationship assumption is the major difference comparing to Pearson correlation, which is built on satisfying the much restrictive linear relationship. Therefore, assume that two variables $X_i$ and $Y_i$ are the original variables

expecting to be evaluated, where $i$ is the paired score and $i \in [1, n]$ as $n$ is the number of observations for each variable. In addition, $x_i$ and $y_i$ are their corresponding ranked values. As a reminder, the Spearman rank correlation $s$ is calculated by the formula listed below:

$$s = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2 - 1)}, \tag{3}$$

where $n$ is the number observations.

Therefore, under the null hypothesis circumstance, the Spearman correlation coefficient can be estimated by the $t$-distribution with degree of freedom of $n - 2$. The $t$ statistics is then calculated by:

$$t = \frac{s\sqrt{n-2}}{\sqrt{1-s^2}}. \tag{4}$$

The Spearman correlation coefficient has the values that satisfy $-1 \leq s \leq 1$, where values $-1$ and 1 refer to perfect monotonic relationship, whilst $s = 0$ indicates monotonically independent random variables. It has been a significant improvement that the Spearman rank correlation coefficient extend the restriction of linearity to monotonic relationship. However, the results of "independent" tendency still cannot reject the possibility of nonlinear association.

#### 2.1.3. *Kendall $\tau$ rank correlation coefficient*

Kendall correlation is proposed in [8] as an updated version of rank correlation measure. It considered the possible differences of ranking orders corresponding to random observers and developed the index $\tau$ to represent the new rank correlation coefficient as below by following [8]:

$$\tau = \frac{actual\ score}{maximum\ possible\ score}, \tag{5}$$

where, in terms of $n$ observations, the *actual score* is the number of different pairs between these two ordered sets, called the symmetric difference distance [9].

Therefore, the *maximum possible score* can be calculated by

$$maximum\ possible\ score = (n-1) + (n-2) + \cdots + 1 = \frac{n(n-1)}{2}. \tag{6}$$

As an alternative rank correlation measure comparing to Spearman rank correlation, the Kendall rank correlation can also detect possible monotonic relationships. As the standard deviation of $\tau$ can be computed by $\sigma_\tau = \frac{1}{3}\sqrt{\frac{2(2n+5)}{n(n-1)}}$, therefore, the null hypothesis test process of obtaining significant test statistics is introduced below by following [9]:

$$Z_\tau = \frac{\tau}{\sigma_\tau}, \tag{7}$$

where $Z_\tau$ is a normal distributed statistics, also satisfies mean of 0 and standard deviation of 1.

## 2.2. *Nonlinear association*

### 2.2.1. *Hoeffding's D test*

Hoeffding's D test is another nonparametric test of independence between two random variables proposed and named after Hoeffding [10]. The major difference between Hoeffding's D test and classical linear association methods like Pearson and Spearman is that it can detect some level of nonlinearity. It is based on ranked value similar as Spearman, however, the difference is that it measures the joint ranked values of two examined variables together.

As a reminder, assume two random variables $X$ and $Y$ with $n$ observations each, in which $x_i$ and $y_i$ have the ranks representing as $RX_i$ and $RY_i$ respectively ($i \in (1, n]$). Additionally, $Q_i$ refers to the number of points with both $x$ and $y$ values less than their corresponding $i$th point. Therefore, $Q_i = \sum_{j=1}^{n} \phi(x_j, x_i)\phi(y_j, y_i)$ and the Hoeffding's D statistic can be calculated as the formula listed below:

$$D = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)}, \tag{8}$$

in which by setting as follows:

$$\begin{aligned} A &= \sum_{i=1}^{n} (RX_i - 1)(RX_i - 2)(RY_i - 1)(RY_i - 2) \\ B &= \sum_{i=1}^{n} (RX_i - 2)(RY_i - 2)Q_i, \quad C = \sum_{i=1}^{n} Q_i(Q_i - 1). \end{aligned} \tag{9}$$

### 2.2.2. *Distance correlation*

Distance correlation is proposed in [11] as a new measure of dependence between random vectors, which also claimed to be designed for detecting nonlinearity. It adopted the empirical concept of Euclidian distance together with sample moments. It is stated in [20] that it is easy to calculate and can be apply to sample sizes $n \geq 2$ without restrictions on matrix inversion or estimation of parameters.

Assume two random variables $X$ and $Y$ with $n$ observations each, for which the pairwise Euclidean distances $a_{ij}$ and $b_{ij}$ (where $i, j = 1, \ldots, n$) can be calculated by:

$$a_{ij} = |x_i - x_j|, \quad b_{ij} = |y_i - y_j|. \tag{10}$$

Therefore, transformed distance matrices $A_{ij}$ and $B_{ij}$ can be defined by:

$$\begin{aligned} A_{ij} &= a_{ij} - \frac{1}{n}\sum_{i=1}^{n} a_{ij} - \frac{1}{n}\sum_{j=1}^{n} a_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \\ B_{ij} &= b_{ij} - \frac{1}{n}\sum_{i=1}^{n} b_{ij} - \frac{1}{n}\sum_{j=1}^{n} b_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}. \end{aligned} \tag{11}$$

Then the distance covariance can be calculated by following

$$V_{xy}^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}, \tag{12}$$

and the distance correlation can be computed by

$$R^2 = \frac{V_{xy}^2}{V_x V_y}, \tag{13}$$

which satisfies $0 \leq R \leq 1$ and is employed to measure the dependency between $X$ and $Y$.

### 2.2.3. *Mutual information*

According to [12], the mutual information are applied to measure the information that two tested variables share with each other, or the same concept that to measure how much knowing one of these variables reduces our uncertainty about the other. The mutual information can be expressed as the following formula in accordance with [12]:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y). \tag{14}$$

where $H(X)$ and $H(Y)$ are the marginal entropies, $H(X|Y)$ and $H(Y|X)$ are the conditional entropies, and $H(X,Y)$ is the joint entropy of $X$ and $Y$. The mutual information defined above takes a value between 0 and infinity, $0 \leq I(X,Y) \leq +\infty$, which makes the comparisons difficult between different samples [12]. In this context, [13] among others, defined and used a standard measure for the mutual information:

$$\lambda = (1 - \exp[-2I(X,Y)])^{\frac{1}{2}}. \tag{15}$$

Note that $\lambda$ captures the overall dependence, both linear and nonlinear, between $X$ and $Y$.

Additionally, in terms of the mutual information of two continuous random variables $X$ and $Y$, it is defined as below [12]:

$$I(X;Y) = \int_X \int_Y P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) d_y d_x, \tag{16}$$

where $P(x,y)$ is the joint probability distribution function of $X$ and $Y$, and $P(x)$ and $P(y)$ are the marginal probability distribution functions of $X$ and $Y$, respectively.

### 2.2.4. *Maximal information coefficient*

According to [14], Maximal Information Coefficient (MIC) is a recently proposed measure of association based on the mutual information which measures that if

a relationship between two random variables exists, a grid can be drawn on the scatter plot of the two variables for partitioning the data points and encapsulating this relationship. The details of definition and calculation of MIC are listed as follows, which we mainly follow [14].

Specifically, for a give finite set $C$ of ordered pairs,the $x$ and $y$ values of $C$ are partitioned into $x$ and $y$ bins respectively, which is defined as an $x - by - y$ grid. As $C|_G$ refers to the distribution induced by the points in $C$ on the cells of $G$, the mutual information of $C|_G$ can be expressed as $I(C|_G)$. Therefore, the MIC of a set $C$ of two variables $X$ and $Y$ with $n$ observations each can be computed by:

$$MIC(X,Y) = \max_{|X||Y|<B} \frac{\max I(C|_G)}{\log(\min(|X||Y|))}, \tag{17}$$

where $|X|$ and $|Y|$ are the number of bins for each variable respectively, and default setting of $B = n^{0.6}$ provides the upper bound of the size of the grids. The MIC measure of association will result in a coefficient in the range of $[0, 1]$, which plays better criterion of association cooperation than mutual information.

## 3. Novel Mutual Association Measure

In this following section, we briefly introduce the adopted advanced techniques and propose a new mutual association measure built on the eigenvalue-based distance with detailed formulation process.

### 3.1. *Singular value decomposition*

Singular Value Decomposition (SVD) is closed related to the Singular Spectrum Analysis (SSA), which is a relatively new, powerful and applicable technique known for both time series analysis and forecasting by wide applications on a range of different fields [12, 15–19]. The SSA technique is performed in two stages, which are known as decomposition and reconstruction. SVD is one of the two significant steps of decomposition after embedding.

The SVD technique adopted in this paper is specifically based on the multivariate extension of SSA (MSSA), where the detailed descriptions of MSSA steps and implementations can be found in [17, 19, 21]. Note that the structures of constructing Hankel matrix containing multiple variables differ by either horizontal or vertical forms. Here in this paper, the following formulation steps of SVD are provided in terms of the vertical form scenario.[a]

Consider $M$ time series with different series length $N_i : Y_{N_i}^{(i)} = (y_1^{(i)}, \ldots, y_{N_i}^{(i)})$ $(i = 1, \ldots, M)$. 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 $[X_1^{(i)}, \ldots, X_{K_i}^{(i)}]$ with vectors $X_j^{(i)}$ that equals to $(y_j^{(i)}, \ldots,$

[a]Note that the horizontal form formulation process are not reproduced here, which can be find with details in [21].

$y_{j+L_i-1}^{(i)})^T \in \mathbf{R}^{L_i}$, where $L_i (2 \leq L_i \leq N_i/2)$ 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 $\mathbf{X}^{(i)} = [X_1^{(i)}, \ldots, X_{K_i}^{(i)}] = (x_{mn})_{m,n=1}^{L_i,K_i}$ after this step. The above procedure for each series separately provides M different $L_i \times K_i$ trajectory matrices $\mathbf{X}^{(i)}$ $(i = 1, \ldots, M)$.

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. Denote $\lambda_{V_1}, \ldots, \lambda_{V_{L_{\text{sum}}}}$ as the eigenvalues of $\mathbf{X}_V \mathbf{X}_V^T$, arranged in decreasing order $(\lambda_{V_1} \geq \cdots \lambda_{V_{L_{\text{sum}}}} \geq 0)$ and $U_{V_1}, \ldots, U_{V_{L_{\text{sum}}}}$, the corresponding eigenvectors, where $L_{\text{sum}} = \sum_{i=1}^{M} L_i$. 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 $Y_{N_i}^{(i)}$, 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 $\mathbf{X}_V = \mathbf{X}_{V_1} + \cdots + \mathbf{X}_{V_{L_{\text{sum}}}}$, 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 $\lambda_{V_i} = 0$).

### 3.2. *Eigenvalue-based distance and novel mutual association measure*

Eigenvalue-based approach is combined with image processing in [16] by considering digital image as matrix of grey level or color values. In which, the authors proposed the relatively new method for image denoising by combining MSSA technique and modified Frobenius distance formula based on eigenvalues. One of their significant research outcomes we adopt here is the concept of eigenvalue-based distances between images. The eigenvalue-based distance for image processing proved with promising performances in image denoising and can be widely applied for face

recognition and verification as another competitive approach [16]. The theoretical formula of eigenvalue-based distance is listed below accordingly, which we mainly follows [16].

Briefly, the eigenvalue-based distance introduced by [16] is built on the trajectory matrices of the images and their SVD expansions. Assume that we have two trajectory matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of size $g \times q$, which are associate with two corresponding images $\mathbf{I}^{(1)}$ and $\mathbf{I}^{(2)}$ of the same size $h \times w$. In order to compare with uniform standard, these two matrices are firstly normalized by the formulation process below:

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{X}^{(1)} \Big/ \sqrt{tr(\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^T)} \\ \mathbf{Y}_2 &= \mathbf{X}^{(2)} \Big/ \sqrt{tr(\mathbf{X}^{(2)}(\mathbf{X}^{(2)})^T)}. \end{aligned} \tag{18}$$

Then the corresponding eigenvalues of matrices $\mathbf{Y}_1\mathbf{Y}_1^T$ and $\mathbf{Y}_2\mathbf{Y}_2^T$ (where both of them are nonnegative definite) can be obtained by SVD, which we use $\lambda_1 \geq \cdots \geq \lambda_g$ and $\mu_1 \geq \cdots \geq \mu_g$ to represent respectively. Note that $tr(\mathbf{Y}_1\mathbf{Y}_1^T) = tr(\mathbf{Y}_2\mathbf{Y}_2^T) = 1$, consequently, for all $i$, $\sum_{i=1}^{g} \lambda_i = \sum_{i=1}^{g} \mu_i = 1$ and corresponding eigenvalues satisfy $\lambda_i \geq 0$, $\mu_i \geq 0$.

A joint trajectory matrix based on $\mathbf{Y} = \binom{\mathbf{Y}_1}{\mathbf{Y}_2}$ is created to analyze two images simultaneously by:

$$\mathbf{Y}\mathbf{Y}^T = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} (\mathbf{Y}_1 \quad \mathbf{Y}_2) = \begin{pmatrix} \mathbf{Y}_1\mathbf{Y}_1^T & \mathbf{Y}_1\mathbf{Y}_2^T \\ \mathbf{Y}_2\mathbf{Y}_1^T & \mathbf{Y}_2\mathbf{Y}_2^T \end{pmatrix}. \tag{19}$$

Consequently, the eigenvalues of the joint trajectory matrix above can be donated as $\upsilon_1 \geq \cdots \geq \upsilon_{2g} \geq 0$, where $\upsilon$ satisfy $\sum_{i=1}^{2g} \upsilon_i = 2$ in accordance of $tr(\mathbf{Y}\mathbf{Y}^T) = tr(\mathbf{Y}_1\mathbf{Y}_1^T) + tr(\mathbf{Y}_2\mathbf{Y}_2^T) = 2$.

As [22] proved that for any matrix of joint form like $\mathbf{Y}\mathbf{Y}^T$, there exists the following relationship between corresponding eigenvalues for any positive integer $k$:

$$\sum_{j=1}^{k} \lambda_j + \sum_{j=1}^{k} \mu_j \geq \sum_{j=1}^{k} \upsilon_j. \tag{20}$$

By defining the cumulative distribution function on the integers $\{1, \ldots, g\}$ or $\{1, \ldots, 2g\}$ respectively, we therefore have

$$F_1(t) = \sum_{j=1}^{[t]} \lambda_j, \quad F_2(t) = \sum_{j=1}^{[t]} \mu_j, \quad F(t) = \frac{1}{2} \sum_{j=1}^{[t]} \upsilon_j \tag{21}$$

which indicate the inequality $F_1(t) + F_2(t) - 2F(t) \geq 0$ for all $t \geq 0$. Finally the distance based on eigenvalue can be formulated as:

$$G(t) = F_1(t) + F_2(t) - 2F(t). \tag{22}$$

More specifically, the natural definition of the eigenvalue-based distance between two images $\mathbf{I}^{(1)}$ and $\mathbf{I}^{(2)}$ can be expressed by:

$$d_1(\mathbf{I}^{(1)}, \mathbf{I}^{(2)}) = \int_0^k G(t)dt = \sum_{j=1}^{k} (\lambda_j + \mu_j - \upsilon_j). \tag{23}$$

The new mutual association measure is then obtained based on the fundamental concept of eigenvalue-based distance. Assume we have two random series $X$ and $Y$, which have the same number of observations $n$. First step, these two random series $X$ and $Y$ are transformed into two dimensional trajectory matrices $\mathbf{M_x}$ and $\mathbf{M_Y}$ by multiplying their transpose series respectively, which can be expressed as:

$$\begin{aligned} \mathbf{M_X} &= XX^T \\ \mathbf{M_Y} &= YY^T. \end{aligned} \tag{24}$$

Considering the concept is built on the relationships between eigenvalues, those two trajectory matrices are normalized before further formulation, which followed the normalization algorithm below:

$$\begin{aligned} \mathbf{NM_X} &= \mathbf{M_X} \Big/ \sqrt{tr(\mathbf{M_X}\mathbf{M_X^T})} \\ \mathbf{NM_Y} &= \mathbf{M_Y} \Big/ \sqrt{tr(\mathbf{M_Y}\mathbf{M_Y^T})}. \end{aligned} \tag{25}$$

Therefore, the joint matrix $\mathbf{NM}$ is created by combining $\mathbf{NM_X}$ and $\mathbf{NM_Y}$:

$$\mathbf{NM} = \begin{pmatrix} \mathbf{NM_X} \\ \mathbf{NM_Y} \end{pmatrix}. \tag{26}$$

Note that in terms of forming this joint matrix, horizontally and vertically formed structures do not have difference for the next step of transforming to trajectory matrices as they will show symmetric feature and provide identical eigenvalues.

The joint matrix $\mathbf{NM}$ then get transformed into trajectory matrix by multiplying its transpose matrix:

$$\mathbf{TM} = \mathbf{NM} \cdot \mathbf{NM^T}, \tag{27}$$

where we use $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_n \geq 0$ to donate the corresponding eigenvalues of $\mathbf{TM}$.

In fact by combining the two random series $X$ and $Y$, the final joint trajectory matrix will provide two significant eigenvalues $\xi_1$ and $\xi_2$, which are the first two in order. The identical (or closely associated) features will be able to presented by the first eigenvalue $\xi_1$ without any information left. In other words, the second eigenvalue $\xi_2$ indicates the information of "distance" between these two series, which also represents the not associated information between these two series. More specifically, if the two random series are identical, the eigenvalues of the joint matrix $\mathbf{TM}$ will show $\xi_1 = 2$ and $\xi_2 = \cdots = \xi_n = 0$. Additionally, on the contrary of the

perfectly identical scenario, meaning if these two series are not associated at all, the corresponding $\xi_1$ and $\xi_2$ of **TM** will be both extremely close or equal to 1.

In summary, we donate $\varphi(X, Y)$ as the mutual association index between two random variables $X$ and $Y$. The definition formula of $\varphi(X, Y)$ is written accordingly as

$$\varphi(X, Y) = 1 - \xi_2, \tag{28}$$

in which $\varphi(X, Y)$ satisfies $0 \leq \varphi(X, Y) \leq 1$. Specifically, $\varphi(X, Y) = 1$ indicates $X$ and $Y$ are most significantly associated (identical); $\varphi(X, Y) = 0$ refers that there is almost no association between $X$ and $Y$.

## 4. Evaluation of Simulations

The performances of both empirical and newly proposed association measures are summarized below by simulations, in which we simulate different representative linear and nonlinear relationships or patterns for investigation and comparison. For each specific relationship (linear or nonlinear), we generate group of series with 200 number of observations for each specific population correlation values and repeat this process 1000 times. All statistics results are summarized and listed for comparison in Table 1, where 2.5% and 97.5% quartile, mean and standard deviation of test statistics from corresponding simulations are provided. Note that all simulations are obtained by R program with corresponding packages, in which representative nonlinear patterns are adopted by referring to the codes in [23].

According to the results in Table 1 by simulations of representative groups of series, the results we obtained are in line with those previous literatures of [14, 24]. In more details, Pearson and Spearman coefficients work properly on simulated linear group as usual with promising results and small standard deviation, while it is noticed that the standard deviation of both Pearson and Spearman correlation coefficients slightly increase when the population correlation coefficients converge to 0; Kendall coefficient provides coefficients with higher variations comparing to corresponding populations and the standard deviations are higher than both Spearman and Pearson with the same slight increasing trend as population coefficients converging to 0; in terms of the simulated nonlinear groups, all three linear measures cannot pick up any relationships, which provide coefficients equal or very close to 0.

In terms of the empirical nonlinear association measures, Hoeffding' s D test, mutual information and MIC cannot provide proper association indices for simulated linear groups comparing to other measures, whilst Distance Correlation is the only one can possibly measure the association by providing relatively closer indices if we take no account of the direction of correlation. These results also confirm the findings in [24] that the Hoeffding's D and MIC appeared to get more differences away from the defined level of population whilst the Distance Correlation got much less. We also notice that for Hoeffding's D test, mutual information and MIC, conversely, their standard deviations show tendency of slight decreasing when

Table 1. Evaluations of association measures by simulated groups of series.

| | Population | New Approach | | | | Pearson | | | | Spearman | | | | Kendall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2.5% | Mean | 97.5% | St.D | 2.5% | Mean | 97.5% | St.D | 2.5% | Mean | 97.5% | St.D | 2.5% | Mean | 97.5% | St.D |
| Linear | 0.8 | 0.8 | 0.8 | 0.8 | 0 | 0.78 | 0.79 | 0.82 | 0.01 | 0.76 | 0.78 | 0.81 | 0.01 | 0.51 | 0.59 | 0.67 | 0.04 |
| | 0.6 | 0.6 | 0.6 | 0.6 | 0 | 0.56 | 0.60 | 0.64 | 0.02 | 0.54 | 0.58 | 0.62 | 0.02 | 0.30 | 0.41 | 0.51 | 0.05 |
| | 0.4 | 0.4 | 0.4 | 0.4 | 0 | 0.34 | 0.39 | 0.45 | 0.03 | 0.33 | 0.38 | 0.44 | 0.03 | 0.14 | 0.26 | 0.37 | 0.06 |
| | 0 | 0 | 0 | 0 | 0 | −0.06 | 0 | 0.06 | 0.03 | −0.06 | 0 | 0.06 | 0.03 | −0.14 | 0 | 0.12 | 0.07 |
| | −0.4 | 0.4 | 0.4 | 0.4 | 0 | −0.45 | −0.40 | −0.34 | 0.03 | −0.44 | −0.38 | −0.33 | 0.03 | −0.38 | −0.26 | −0.14 | 0.06 |
| | −0.6 | 0.6 | 0.6 | 0.6 | 0 | −0.64 | −0.60 | −0.56 | 0.02 | −0.62 | −0.58 | −0.54 | 0.02 | −0.51 | −0.41 | −0.30 | 0.06 |
| | −0.8 | 0.8 | 0.8 | 0.8 | 0 | −0.82 | −0.79 | −0.77 | 0.01 | −0.81 | −0.79 | −0.76 | 0.01 | −0.67 | −0.59 | −0.51 | 0.04 |
| Non-linear | wave | 0.01 | 0.07 | 0.16 | 0.04 | −0.06 | 0 | 0.06 | 0.03 | −0.06 | 0 | 0.07 | 0.03 | −0.16 | 0 | 0.17 | 0.08 |
| | trapezoid | 0.52 | 0.63 | 0.71 | 0.05 | −0.06 | 0 | 0.05 | 0.03 | −0.06 | 0 | 0.05 | 0.03 | −0.11 | 0 | 0.10 | 0.06 |
| | diamond | 0.01 | 0.05 | 0.14 | 0.04 | −0.04 | 0 | 0.04 | 0.02 | −0.05 | 0 | 0.05 | 0.02 | −0.11 | 0 | 0.12 | 0.06 |
| | quadratic | 0.01 | 0.12 | 0.33 | 0.09 | −0.07 | 0 | 0.08 | 0.04 | −0.08 | 0 | 0.08 | 0.04 | −0.20 | 0 | 0.19 | 0.11 |
| | cross | 0.01 | 0.12 | 0.33 | 0.09 | −0.09 | 0 | 0.08 | 0.05 | −0.08 | 0 | 0.07 | 0.04 | −0.21 | 0 | 0.23 | 0.12 |
| | circle | 0.01 | 0.08 | 0.23 | 0.06 | −0.04 | 0 | 0.04 | 0.02 | −0.03 | 0 | 0.03 | 0.01 | −0.14 | 0 | 0.13 | 0.07 |
| | cluster | 0.01 | 0.04 | 0.10 | 0.03 | −0.02 | 0 | 0.01 | 0.01 | −0.05 | 0 | 0.04 | 0.02 | −0.09 | 0 | 0.09 | 0.05 |
| | Population | Hoeffding's D Test | | | | Distance Correlation | | | | Mutual Information | | | | MIC | | | |
| | | 2.5% | Mean | 97.5% | St.D | 2.5% | Mean | 97.5% | St.D | 2.5% | Mean | 97.5% | St.D | 2.5% | Mean | 97.5% | St.D |
| Linear | 0.8 | 0.23 | 0.26 | 0.28 | 0.02 | 0.73 | 0.75 | 0.78 | 0.01 | 0.30 | 0.39 | 0.48 | 0.05 | 0.47 | 0.51 | 0.56 | 0.03 |
| | 0.6 | 0.09 | 0.11 | 0.14 | 0.01 | 0.51 | 0.55 | 0.59 | 0.02 | 0.14 | 0.21 | 0.28 | 0.04 | 0.28 | 0.32 | 0.37 | 0.02 |
| | 0.4 | 0.03 | 0.04 | 0.06 | 0.01 | 0.31 | 0.36 | 0.42 | 0.03 | 0.06 | 0.11 | 0.17 | 0.03 | 0.18 | 0.21 | 0.24 | 0.02 |
| | 0 | 0 | 0 | 0 | 0 | 0.04 | 0.06 | 0.08 | 0.01 | 0.01 | 0.04 | 0.09 | 0.02 | 0.12 | 0.13 | 0.15 | 0.01 |
| | −0.4 | 0.03 | 0.04 | 0.06 | 0 | 0.31 | 0.36 | 0.41 | 0.03 | 0.06 | 0.11 | 0.17 | 0.03 | 0.18 | 0.21 | 0.24 | 0.02 |
| | −0.6 | 0.09 | 0.11 | 0.14 | 0.01 | 0.51 | 0.55 | 0.60 | 0.02 | 0.13 | 0.21 | 0.29 | 0.04 | 0.28 | 0.32 | 0.36 | 0.02 |
| | −0.8 | 0.23 | 0.25 | 0.28 | 0.01 | 0.73 | 0.76 | 0.78 | 0.01 | 0.29 | 0.39 | 0.49 | 0.05 | 0.47 | 0.51 | 0.56 | 0.02 |
| Non-linear | wave | 0.01 | 0.01 | 0.02 | 0 | 0.33 | 0.40 | 0.48 | 0.04 | 0.22 | 0.37 | 0.51 | 0.08 | 0.96 | 0.99 | 1.00 | 0.01 |
| | trapezoid | 0 | 0 | 0 | 0 | 0.15 | 0.20 | 0.27 | 0.03 | 0.03 | 0.08 | 0.14 | 0.03 | 0.17 | 0.19 | 0.22 | 0.01 |
| | diamond | 0 | 0 | 0 | 0 | 0.15 | 0.20 | 0.28 | 0.03 | 0.06 | 0.12 | 0.19 | 0.04 | 0.13 | 0.15 | 0.17 | 0.01 |
| | quadratic | 0.09 | 0.10 | 0.11 | 0 | 0.46 | 0.51 | 0.56 | 0.03 | 0.71 | 0.78 | 0.86 | 0.04 | 0.64 | 0.69 | 0.74 | 0.03 |
| | cross | 0.04 | 0.05 | 0.05 | 0 | 0.30 | 0.36 | 0.45 | 0.04 | 0.59 | 0.76 | 0.91 | 0.09 | 0.55 | 0.57 | 0.58 | 0.01 |
| | circle | 0.03 | 0.04 | 0.05 | 0 | 0.12 | 0.17 | 0.26 | 0.04 | 0.02 | 0.04 | 0.07 | 0.02 | 0.55 | 0.56 | 0.57 | 0.01 |
| | cluster | 0 | 0 | 0 | 0 | 0.07 | 0.10 | 0.14 | 0.02 | 1.38 | 1.38 | 1.38 | 0 | 0.12 | 0.13 | 0.14 | 0.01 |

the population correlation coefficients converge to 0, which may indicate that these association measures do detect some level of linear relationship, while the consistency and accuracy level are not stable as the other measures. Regarding the results of simulated nonlinear groups: Hoeffding' s D test shows very limited capacity of detecting any possible relationships; Distance Correlation is relatively more sensitive on nonlinear patterns; mutual information and MIC detect different levels of association for different linear patterns, in which, mutual information shows relatively significant estimates for quadratic and cross patterns, additionally, due to its algorithm of discrediting data for calculation, it give same value for the cluster pattern, whilst MIC gives significant estimates for wave and provides relatively significant indices for quadratic, cross and circle patterns.

Considering the performance of this novel mutual association measure by eigenvalue-based criterion, it is worth to be noted that for the simulated linear group, the mutual association measure by eigenvalue-based distance achieve solid and consistent indices, which are precisely identical to the absolute values of corresponding generated population correlation coefficients. As a mutual association measure built on the eigenvalue-based distance, the brief concept of this measure is identifying the information shifted to the second eigenvalue of a matrix formed by a multivariate system. Therefore, the mutual association it can detect do not consider the direction of effect. Actually, in general, the direction of effect can easily be noticed by a simple time series diagram. Comparing to the other widely accepted association measures, only mutual association measure by eigenvalue-based distance provides almost permanently stable results for 1000 times of simulations.

Regarding the performance of mutual association on nonlinear patterns, it is noticed that only trapezoid pattern can be detected with relatively significant statistics. It cannot provide more significant evidences for other nonlinear patterns, whilst considering the other empirical linear association measures, the novel mutual association measure is slightly more sensitive than Hoeffding's D Test, with fairly less significant results for quadratic and cross. Moreover, in terms of the trapezoid pattern, it has not been significantly detected by any other listed measures.

In general, according to the evaluations of all listed association measures, there is no measure that can well perform for detecting both linear and nonlinear relationships with also relatively accurate estimates. The highlight point for novel mutual association measure is that it shows consistent and precise estimates for all linear simulations with 0 variation, and it can detect the trapezoid pattern with significant estimates that all previously listed measure could not achieve.

## 5. Performances in Real Data of Oil Price and Tourist Arrivals

Considering the complexity of real data and the restricted nonlinear relationships simulations can offer for evaluation, here in this section we considered one case of real data for further investigation and comparison. Note that all preconditions of each measure are satisfied respectively. It is also worth to be highlighted that we

are not making any assumptions or models on data that are undertaking tests as the aim of this paper is proposing a novel association measure and evaluating the performances by comparing to empirical linear and nonlinear association measures from the statistical data analysis point of view.

The data used are at monthly frequency covering the period from January 1996 to December 2015 of UK. In terms of the data, UK tourist arrivals were obtained from the Eurostat. Data of Europe Brent Spot Price (BRT) in the unit of dollars per barrel is obtained from the EIA [29]. Figure 1 shows the time series plot of the monthly oil prices, whilst, Fig. 2 presents the time series plots of the monthly tourist arrivals. It can be observed that the tourist arrivals for UK clearly shows significant feature of cycle with possible existing trend.

The emerging concerns of oil price and its impacts on diverse aspects of economy have been studies by numerous researchers recent decades with well established scientific literatures [27]. Among which, the relationship between oil price and tourism has drawn significant attentions. A critical review of the studies of tourism and oil can be found in [28] for reference. Table 2 summarizes the results of both empirical and novel association measures adopted. It is observed that the empirical methods cannot possibly detect any association whilst the novel measure achieves significant evidences.

More specifically, the results are very similar between the coefficients of Pearson and Spearman, which both show relatively low levels of association less than 0.3. Kendall correlation, Distance Correlation and MIC reflect similar levels of significance, whilst Hoeffding's D Test has the lowest level of sensitivity. The novel mutual

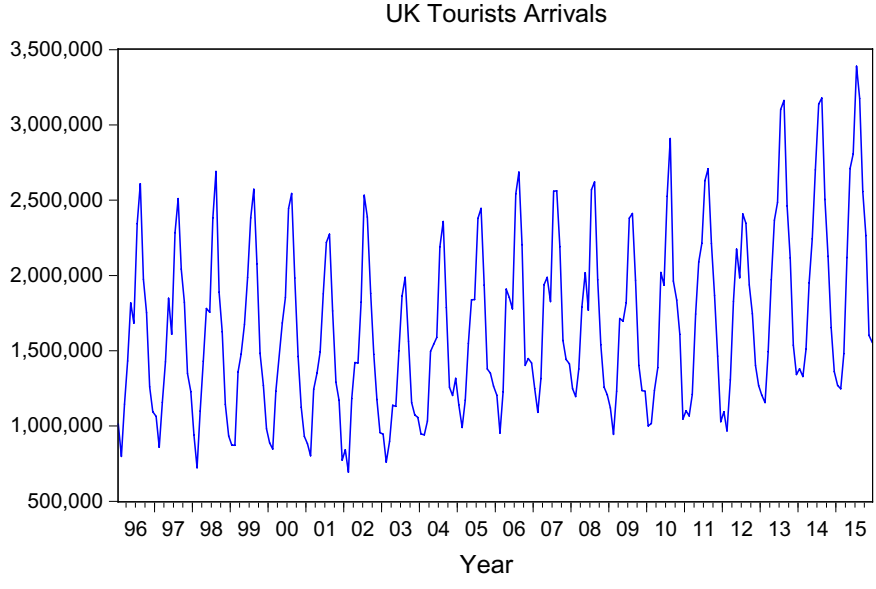


Fig. 1. Monthly oil price data from 1996 to 2015.

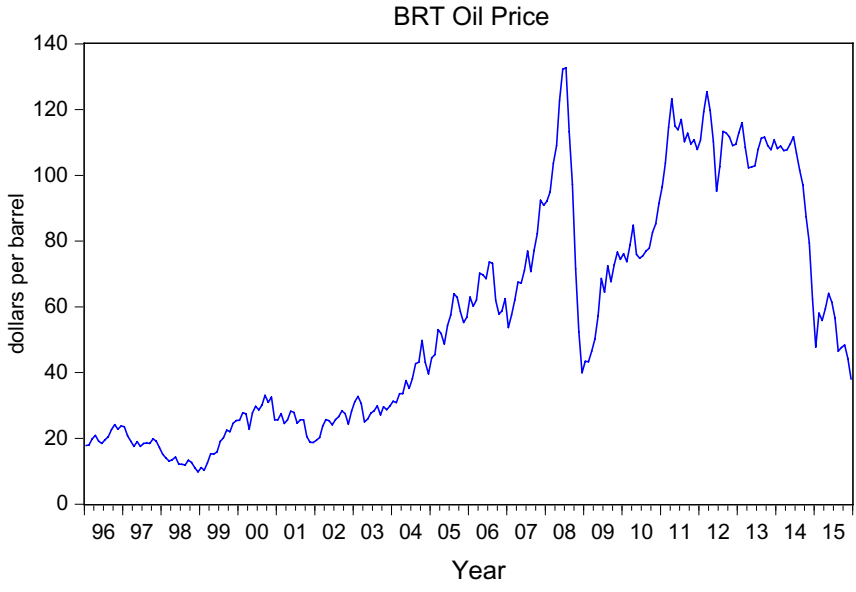


Fig. 2. Monthly tourists arrivals of UK from 1996 to 2015.

Table 2. Comparison of association measures on analyses of oil prices and tourist arrivals in UK.

| Country | Association Measures | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | New Approach | Pearson | Spearman | Kendall | HoefD | DisCorr | MI | MIC |
| UK | **0.85** | 0.29** | 0.28** | 0.19** | 0.02** | 0.27 | 0.07 | 0.28 |

*Note*: ** indicates the significance of 5%.

association measure, on the other hand, can detect the possible association and provide significant result. The novel mutual association measure greatly outperforms the empirical methods and indicates the significant mutual association between tourist arrivals and oil price with evidences for UK. The initially adopted novel mutual association measure successfully proves the advantage on nonlinear association detection in complex system like oil-tourism studies. It is sensitive enough to confirm the crucial relationship between the tourist arrivals and oil prices by relatively less amount of data to contribute to the study of a complex economy system.

## 6. Conclusion

Considering the crucial importance of association study in better understanding multivariate systems across various disciplines, this paper aims at proposing a novel mutual association measure by combing the eigenvalue-based technique. The performance of the novel association measure is evaluated with comparisons by

simulations as well as the case of real data. Moreover, the performances achieved are significantly promising and it has to be highlighted that its valuable potentials on nonlinear association studies in complex systems across numerous subjects.

This paper is the first attempt of employing eigenvalue-based technique with a multivariate system into the development of an alternative or better performed association measure. In terms of the evaluations by simulated linear and nonlinear relationships, it currently may not master on identifying all nonlinear patterns, whilst it gives highest significant reaction for trapezoid nonlinear pattern without losing the ability on linear association detection that other measures cannot achieve. Considering a real data case of oil prices and tourist arrivals in UK, it is significantly evidenced that the novelly developed association measure is a reliable, sensitive, assumption free approach that can outperform or at least being alternative method comparing to empirical measures in the study of a complex economy system.

In general, there should not be a restriction of one specific association measure as the advantages of different measures vary significantly, which give more options for the association analysis of random groups of series in a complex system like economics and social science. However, this paper obtains solid evidences from both simulations and real case that this novel method can identify complex associations which empirical methods may fail to detect. The advantage of this method is that it does not restrict on the domain of either linearity or nonlinearity, but consider the associated information of the series as a whole. Additionally, the calculations are efficient and convenient. In summary, this paper is the temporary summary about the beginning of this development. There will be many possibilities for further improvements as the next stage of this study, for example, expanding the nonlinear associations or combinations of linearity and nonlinearity for simulated evaluation, developing the direction of association into the current index outcome, etc.

## References

[1] Filis, G., Degiannakis, S. and Floros, C. (2011). Dynamic correlation between stock market and oil prices: The case of oil-importing and oil-exporting countries. *International Review of Financial Analysis*, **20**(3), 152–164.

[2] Huang, X. (2015). A comparison of association measures: Evidence from stock markets and oil prices. *International Journal of Energy and Statistics*, **3**(03), 1550012.

[3] Hajian, S. and Movahed, M. S. (2010). Multifractal detrended cross-correlation analysis of sunspot numbers and river flow fluctuations. *Physica A*: *Statistical Mechanics and Its Applications*, **389**(21), 4942–4957.

[4] Chapman, N. (2012). Correlation analysis in chemistry: Recent advances. Springer Science & Business Media.

[5] George, K. W., Chen, A., Jain, A., Batth, T. S., Baidoo, E. E., Wang, G. and Lee, T. S. (2014). Correlation analysis of targeted proteins and metabolites to assess and engineer microbial isopentenol production. *Biotechnology and Bioengineering*, **111**(8), 1648–1658.

[6] Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, **58**(347–352), 240–242.

[7] Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, **15**(2), 201–292.

[8] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1–2), 81–93.

[9] Abdi, H. (2007). The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics*, 508–510.

[10] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19**(3), 293–325.

[11] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **35**(6), 2769–2794.

[12] Hassani, H., Dionisio, A. and Ghodsi, M. (2010). The effect of noise reduction in measuring the linear and nonlinear dependency of financial markets. *Nonlinear Analysis: Real World Applications*, **11**(1), 492–502.

[13] Dionísio, A., Menezes, R. and Mendes, D. A. (2004). Mutual information: A measure of dependency for nonlinear time series. *Physica A: Statistical Mechanics and Its Applications*, **344**(1), 326–329.

[14] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, **334**(6062), 1518–1524.

[15] Hassani, H. (2007). Singular spectrum analysis: methodology and comparison. *Journal of Data Science*, **5**(2), 239–257.

[16] Rodrıguez-Aragón, L. J. and Zhigljavsky, A. (2010). Singular spectrum analysis for image processing. *Statistics and Its Interface*, **3**(3), 419–426.

[17] Hassani, H., Heravi, S. and Zhigljavsky, A. (2013). Forecasting uk industrial production with multivariate singular spectrum analysis. *Journal of Forecasting*, **32**(5), 395–408.

[18] Hassani, H., Webster, A., Silva, E. S. and Heravi, S. (2015). Forecasting US tourist arrivals using optimal singular spectrum analysis. *Tourism Management*, **46**, 322–335.

[19] Hassani, H., Huang, X., Gupta, R. and Ghodsi, M. (2016) Does sunspot numbers cause global temperatures? A reconsideration using non-parametric causality tests. *Physica A: Statistical Mechanics and Its Applications*, **460**, 54–65. Elsevier [DOI:10.1016/j.physa.2016.04.013].

[20] Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, **3**(4), 1236–1265.

[21] Sanei, S. and Hassani, H. (2015). Singular spectrum analysis of biomedical signals. CRC Press.

[22] Thompson, R. C. and Therianos, S. (1972). The eigenvalues of complementary principal submatrices of a positive definite matrix. *Canad. J. Math*, **24**(4), 658–667.

[23] Boigelot, D. (2011). An example of the correlation of $x$ and $y$ for various distributions of $(x, y)$ pairs. *Original Online Commons of Correlation Examples* [Available at: en.wikipedia.org/wiki/File:Correlation_examples2.svg].

[24] Clark, M. (2013). A comparison of correlation measures. *Center for Social Research*, University of Notre Dame.

[25] FRED (2015). Technical report, Federal Reserve Economic Data. https://research.stlouisfed.org/fred2/.

[26] CEIC (2015). Technical report, CEIC Database. http://www.ceicdata.com/en.

[27] Hamilton, J. D. (1996). This is what happened to the oil price-macroeconomy relationship. *Journal of Monetary Economics*, **38**(2), 215–220.
[28] Becken, S. (2011). A critical review of tourism and oil. *Annals of Tourism Research*, **38**(2), 359–379.
[29] EIA. (2016). U.S. Energy Information Administration [Available at:http://www.eia.gov/].