Physica A 414 (2014) 274–284

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A new correlation coefficient for bivariate time-series data

Orhan Erdem [a,1], Elvan Ceyhan [b,2], Yusuf Varli [a,*]

[a] Research Department, Borsa İstanbul, Resitpasa Mahallesi, Tuncay Artun Caddesi, Emirgan, 34467 Istanbul, Turkey
[b] Department of Mathematics, Koç University, 34450 Sariyer, Istanbul, Turkey

HIGHLIGHTS

- We introduce a new correlation coefficient taking the lag difference of data points.
- We investigate the properties of this new correlation coefficient.
- New correlation coefficient captures the cross-independence of two variables over time.
- New coefficient is compared with the Pearson and DCCA coefficients via simulations.

ARTICLE INFO

Article history:
Received 19 March 2014
Received in revised form 2 July 2014
Available online 24 July 2014

Keywords:
Cross-correlation
Pearson's correlation coefficient
DCCA
Stationarity
Non-stationary time series

ABSTRACT

The correlation in time series has received considerable attention in the literature. Its use has attained an important role in the social sciences and finance. For example, pair trading in finance is concerned with the correlation between stock prices, returns, etc. In general, Pearson's correlation coefficient is employed in these areas although it has many underlying assumptions which restrict its use. Here, we introduce a new correlation coefficient which takes into account the lag difference of data points. We investigate the properties of this new correlation coefficient. We demonstrate that it is more appropriate for showing the direction of the covariation of the two variables over time. We also compare the performance of the new correlation coefficient with Pearson's correlation coefficient and Detrended Cross-Correlation Analysis (DCCA) via simulated examples.
© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Various financial models (such as pairs trading) are concerned with the correlation between two different time-series data, e.g., stock prices or returns. Pearson's product moment correlation coefficient is the most commonly used estimator in measuring such correlations. However, there are many underlying assumptions (such as stationarity) for the validity of this coefficient [1].

If a sample set of time-series data is stationary, then the population's mean, variance, and covariance between any two different dates can be estimated based on the sample. If a data is nonstationary, then it violates certain assumptions while estimating these parameters. In general, price series are assumed to be non-stationary, whereas returns are assumed to be stationary. Thus, using Pearson's formula for the calculation of correlation between two price series is not appropriate [2].

* Corresponding author. Tel.: +90 212 298 21 23; fax: +90 212 298 25 00.
E-mail addresses: Orhan.Erdem@borsaistanbul.com (O. Erdem), elceyhan@ku.edu.tr (E. Ceyhan), yusuf.varli@borsaistanbul.com (Y. Varli).
1 Tel.: +90 212 298 22 20; fax: +90 212 298 25 00.
2 Tel.: +90 212 338 18 45; fax: +90 212 338 15 59.

http://dx.doi.org/10.1016/j.physa.2014.07.054
0378-4371/© 2014 Elsevier B.V. All rights reserved.

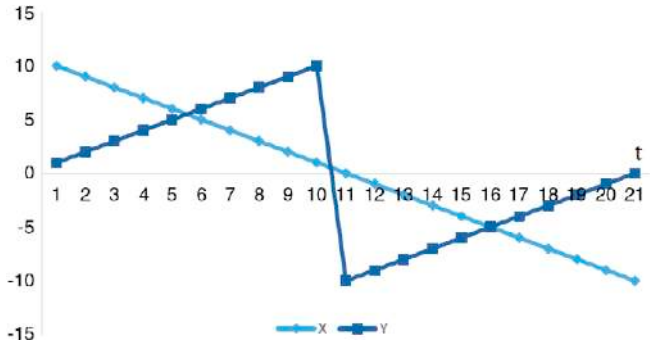

Fig. 1. An illustration of directionality detection problem. Note: $\begin{cases} X_t \text{ points (diamond) are generated as } X_t = 11 - t \text{ and} \\ Y_t \text{ points (square) are generated as } Y_t = t - 1 \text{ for } t = 1 \text{ to11 and } Y_t = t - 22 \text{ for } t = 12, \\ \quad \text{to 21} \end{cases}$

Apart from stationarity, there is another drawback about Pearson's correlation coefficient: it is concerned with the distance of two variables from their means. Assuming that we are interested in two variables that move in the opposite direction while at the same time being both above or below their means. If both of the variables are above (or below) their means, the sum of multiplication of the two variables' deviations from their means will positively contribute to the numerator in Pearson's formula and hence to the correlation coefficient, although the variables move in the opposite direction. The following example (Fig. 1) may illustrate this problem:

In this example, the means of both of the variables is 0, and they are above their means between $t = 0$ and $t = 11$, below their means between $t = 12$ and $t = 21$. Although the variables are moving in the opposite direction almost all the time, Pearson's correlation coefficient, denoted as ρ_p, is 0.50.

A similar idea holds true when two variables move in the same direction while one of the variables is above its mean whereas the other variable is below its mean. In this case, the sum of multiplication of the two variables' deviations from their means will negatively contribute to the numerator in Pearson's formula and hence to the correlation coefficient, even though the variables move in the same direction.

In this article we propose a new correlation coefficient that measures the distance between two subsequent data points by taking the lag difference into consideration. Although the very first data point is lost, we demonstrate that the new correlation coefficient better captures the direction of the covariation of the two variables over time. We also propose various extensions of this coefficient in order to obtain more reasonable and reliable results at the expense of having more complex formulas.

The paper proceeds as follows: we present preliminaries in Section 2. In Section 3, we introduce the new correlation coefficient and discuss its properties. We exhibit a series of simulations to show the characteristics of the new correlation coefficient in Section 4. Furthermore, we conclude our work and point to prospective research directions in Section 5. Finally, we present the matrix forms of correlation coefficients in the Appendix.

2. Preliminaries

Let $P_{i,t}$ (hereafter P_{it} for today, $P_{i,t-s}$ for a lagged time of s units) represents the price of asset i at time t. We will denote the entire sequence of values $\{P_{i1}, P_{i2}, \ldots, P_{iT}\}$ as $\{P_{it}\}$.

The simple return of asset i at time t is defined as:

$$R_{it} = \frac{P_{it} - P_{i,t-1}}{P_{i,t-1}}. \tag{1}$$

Similarly log-return is defined as:

$$r_{it} = \log\left(P_{it}/P_{i,t-1}\right). \tag{2}$$

Let $\{X_t\}$ be a kind of stochastic process; we define the stationarity as follows: a stochastic process $\{X_t\}$ having a finite mean and variance is said to be stationary, if for all t and $t - s$:

$$\mathbf{E}(X_t) = \mathbf{E}(X_{t-s}) = \mu \tag{3}$$

$$\mathbf{E}\left[(X_t - \mu)^2\right] = \mathbf{E}\left[(X_{t-s} - \mu)^2\right] = \sigma^2 \tag{4}$$

$$\mathbf{E}\left[(X_t - \mu)(X_{t-s} - \mu)\right] = \mathbf{E}\left[\left(X_{t-j} - \mu\right)\left(X_{t-j-s} - \mu\right)\right] = \gamma_{(s)} \tag{5}$$

where $\mu, \sigma^2, \gamma_{(s)}$ are all constants i.e., independent of time [3].

In practice, stock prices may be assumed as non-stationary, whereas simple and log returns may be assumed to be stationary [2]. Furthermore, conventionally logarithm of stock prices are assumed to follow *Geometric Brownian Motion* [4] which means that log-returns are assumed to be normally distributed.

3. The new correlation coefficient and its properties

Since Pearson's correlation coefficient, ρ_P, might have some problems for time-series data, we suggest the following correlation coefficient as an alternative:

$$\rho_O = \frac{\alpha_{xy}}{\alpha_x \alpha_y} \tag{6}$$

where $\alpha_x^2 = \mathbf{E}\left[(X_t - X_{t-1})^2\right]$, $\alpha_y^2 = \mathbf{E}\left[(Y_t - Y_{t-1})^2\right]$ and $\alpha_{xy} = \mathbf{E}\left[(X_t - X_{t-1})(Y_t - Y_{t-1})\right]$.

The new correlation coefficient ρ_O can then be estimated as:

$$\widehat{\rho}_O = \frac{A_{xy}}{A_x A_y} \tag{7}$$

where $A_x^2 = \frac{1}{(T-1)}\sum_{t=2}^{T}(X_t - X_{t-1})^2$, $A_y^2 = \frac{1}{(T-1)}\sum_{t=2}^{T}(Y_t - Y_{t-1})^2$ and, $A_{xy} = \frac{1}{(T-1)}\sum_{t=2}^{T}(X_t - X_{t-1})(Y_t - Y_{t-1})$ which are related to the first order autocorrelations of X_t and Y_t, respectively [5]. It should also be noticed that α_{xy} is estimated by A_{xy} and α_x, α_y by A_x, A_y, respectively. For any two processes $\{X_t\}$ and $\{Y_t\}$, it is easy to see that $\mathbf{E}(A_x^2) = \alpha_x^2$, $\mathbf{E}(A_y^2) = \alpha_y^2$, and $\mathbf{E}(A_{xy}) = \alpha_{xy}$.

If we let $X_t = \log(P_{1t})$ for the first stock price and $Y_t = \log(P_{2t})$ for the second stock price, then Eq. (7) can be written as:

$$\widehat{\rho}_O = \frac{\sum_{t=2}^{T} r_{1t} r_{2t}}{\sqrt{\sum_{t=2}^{T} r_{1t}^2 \sum_{t=2}^{T} r_{2t}^2}} \tag{8}$$

where r_{1t}, r_{2t} are defined in Eq. (2).

3.1. Basic properties of the new correlation coefficient

It can be shown that ρ_O satisfies the inequality $|\rho_O| \leq 1$. Using the Cauchy–Schwarz inequality, one can write that the random variables (processes) X_t, Y_t satisfy:

$$\{\mathbf{E}\left[(X_t - X_{t-1})(Y_t - Y_{t-1})\right]\}^2 \leq \mathbf{E}\left[(X_t - X_{t-1})^2\right]\mathbf{E}\left[(Y_t - Y_{t-1})^2\right] \tag{9}$$

with equality holding if and only if one of the variables is almost surely a multiple of the other, that is, $P\left[a(X_t - X_{t-1}) = b(Y_t - Y_{t-1})\right] = 1$ for some real a and b, at least one of which is non-zero.

We can rewrite Eq. (9) as:

$$-1 \leq \frac{\mathbf{E}\left[(X_t - X_{t-1})(Y_t - Y_{t-1})\right]}{\sqrt{\mathbf{E}\left[(X_t - X_{t-1})^2\right]\mathbf{E}\left[(Y_t - Y_{t-1})^2\right]}} = \rho_O \leq 1 \tag{10}$$

since $\mathbf{E}\left[(X_t - X_{t-1})^2\right]$ and $\mathbf{E}\left[(Y_t - Y_{t-1})^2\right]$ are strictly positive.

3.2. The properties under stationarity

Next, we derive the quantities α_x^2, α_y^2, and α_{xy} under the bivariate stationarity assumption for $\{X_t, Y_t\}$, that is, under the assumption that:

$$\begin{aligned}
&\mathbf{E}(X_t) = \mathbf{E}(X_{t-s}) = \mu_X, && \text{(i)}\\
&\mathbf{E}(Y_t) = \mathbf{E}(Y_{t-s}) = \mu_Y,\\
&\mathbf{E}\left[(X_t - \mu_X)^2\right] = \mathbf{E}\left[(X_{t-s} - \mu_X)^2\right] = \sigma_X^2, && \text{(ii)}\\
&\mathbf{E}\left[(Y_t - \mu_Y)^2\right]\mathbf{E}\left[(Y_{t-s} - \mu_Y)^2\right] = \sigma_Y^2,\\
&\mathbf{E}\left[(X_t - \mu_X)(X_{t-s} - \mu_X)\right] = \gamma_X(s), && \text{(iii)}\\
&\mathbf{E}\left[(Y_t - \mu_Y)(Y_{t-s} - \mu_Y)\right] = \gamma_Y(s).
\end{aligned}$$

Furthermore, the cross-covariance between $\{X_t\}$ and $\{Y_t\}$ at lag s [5] is

$$\mathbf{E}\left[(X_t - \mu_X)(Y_{t-s} - \mu_Y)\right] = \gamma_{XY}(s). \tag{iv}$$

Notice that at lag 0 (i.e., $s = 0$), $\gamma_X(0) = \sigma_X^2$, $\gamma_Y(0) = \sigma_Y^2$, and $\gamma_{XY}(0) = \sigma_{XY}^2 = \mathbf{Cov}(X, Y)$.

Then,

$$\begin{aligned}\alpha_x^2 &= \mathbf{E}\left[(X_t - X_{t-1})^2\right] = \mathbf{E}\left[X_t^2 - 2X_tX_{t-1} + X_{t-1}^2\right]\\ &= \mathbf{E}\left[X_t^2\right] - 2\mathbf{E}\left[X_tX_{t-1}\right] + \mathbf{E}\left[X_{t-1}^2\right]\\ &= 2\mathbf{E}\left[X_t^2\right] - 2\mathbf{E}\left[X_tX_{t-1}\right]\\ &= 2(\sigma_X^2 + \mu_X^2) - 2(\gamma_X(1) + \mu_X^2)\\ &= 2(\sigma_X^2 - \gamma_X(1))\end{aligned} \tag{11}$$

since $\mathbf{E}\left[X_t^2\right] = (\sigma_X^2 + \mu_X^2)$ and $\mathbf{E}\left[X_tX_{t-1}\right] = \gamma_X(1) + \mu_X^2$. Similarly, $\alpha_y^2 = 2\left(\sigma_Y^2 - \gamma_Y(1)\right)$.

Furthermore,

$$\begin{aligned}\alpha_{xy} &= \mathbf{E}\left[(X_t - X_{t-1})\left(Y_t - Y_{t-1}\right)\right]\\ &= \mathbf{E}\left[X_tY_t\right] - \mathbf{E}\left[X_tY_{t-1}\right] - \mathbf{E}\left[X_{t-1}Y_t\right] + \mathbf{E}[X_{t-1}Y_{t-1}]\\ &= 2\mathbf{E}[X_tY_t] - 2\mathbf{E}[X_tY_{t-1}]\\ &= 2(\gamma_{XY}(0) + \mu_X\mu_Y) - 2(\gamma_{XY}(1) + \mu_X\mu_Y)\\ &= 2(\gamma_{XY}(0) - \gamma_{XY}(1))\end{aligned} \tag{12}$$

since $\mathbf{E}[X_tY_t] = \gamma_{XY}(0) + \mu_X\mu_Y$ and $\mathbf{E}[X_tY_{t-1}] = \gamma_{XY}(1) + \mu_X\mu_Y$.

But, $\gamma_{XY}(0) = \sigma_{XY}^2$ and $\sigma_{XY}^2 = \rho_{XY}\sigma_X\sigma_Y$ where $\rho_{XY} = \mathbf{Corr}(X, Y)$ which equals ρ_P. Then,

$$\alpha_{xy} = 2(\rho_P\sigma_X\sigma_Y - \gamma_{XY}(1)).$$

Thus, for stationary bivariate time series $\{X_t, Y_t\}$,

$$\rho_O = \frac{\rho_P\sigma_X\sigma_Y - \gamma_{XY}(1)}{\sqrt{\left(\sigma_X^2 + \gamma_X(1)\right)\left(\sigma_Y^2 + \gamma_Y(1)\right)}}. \tag{13}$$

Observe that when $\gamma_X(1) = \gamma_Y(1) = \gamma_{XY}(1) = 0$,

$$\rho_O = \frac{\rho_P\sigma_X\sigma_Y}{\sigma_X\sigma_Y} = \rho_P.$$

Hence, ρ_O and ρ_P agree whenever autocorrelations[3] of $\{X_t\}$ and $\{Y_t\}$ at lag 1 and cross-correlation of $\{X_t\}$ and $\{Y_t\}$ at lag 1 are all zero. Furthermore, under the bivariate stationarity, ρ_P ignores the autocorrelations and cross-correlation, whereas ρ_O incorporates lag 1 autocorrelations.

If $\{X_t\}$ and $\{Y_t\}$ are each (univariate) stationary as in Section 2, but $\{X_t\}$ and $\{Y_t\}$ are independent that is, above (i)–(iii) hold and (iv) holds with $\gamma_{XY}(s) = 0$ for all $s \geq 0$, then $\rho_P = 0$; hence

$$\rho_O = \frac{\rho_P\sigma_X\sigma_Y}{\sqrt{(\sigma_X^2 + \gamma_X(1))(\sigma_Y^2 + \gamma_Y(1))}} = 0. \tag{14}$$

If X_t are i.i.d. with $\mathbf{E}(X_t) = \mu_X$, $\mathbf{Var}(X_t) = \sigma_X^2$ and Y_t are i.i.d. with $\mathbf{E}(Y_t) = \mu_Y$, $\mathbf{Var}(Y_t) = \sigma_Y^2$, then

$$\rho_O = \frac{\rho_P\sigma_X\sigma_Y - \gamma_{XY}(1)}{\sigma_X\sigma_Y} = \rho_P - \frac{\gamma_{XY}(1)}{\sigma_X\sigma_Y} \tag{15}$$

since $\gamma_X(1) = \gamma_Y(1) = 0$ for i.i.d $\{X_t\}$ and $\{Y_t\}$.

If $\{X_t\}$ and $\{Y_t\}$ are both i.i.d. as above and are independent of each other then $\rho_O = \rho_P = 0$, since $\gamma_{XY}(1) = 0$ for independent $\{X_t\}$ and $\{Y_t\}$.

[3] Here, the term of autocorrelation is used based on the definition of Pearson's correlation coefficient. Since first lagged variable is in the formula of the new correlation coefficient, we do not prefer to apply the definition of the new correlation coefficient to measure autocorrelation especially at lag 1. For any i.i.d. X_t, autocorrelation of X_t at lag $n \geq 2$ with respect to new correlation coefficient is $\rho_O(n) = 0$.

Therefore, we once again state that ρ_O and ρ_P measure related but different things. Specifically, they both measure the covariation of $\{X_t\}$ and $\{Y_t\}$ but with emphasis on different aspects. For example, in a hypothesis testing framework, we have $\rho_O = \rho_P = 0$ under independence of $\{X_t\}$ and $\{Y_t\}$. However, if $\{X_t\}$ and $\{Y_t\}$ are independent only at lag 0, but dependent at lag 1 (i.e., $\gamma_{XY}(s) \neq 0$ for $s = 1$), then $\rho_P = 0$ but $\rho_O = \frac{-\gamma_{XY}(1)}{\sigma_X \sigma_Y}$.

3.3. The properties under non-stationarity

In this section, we consider special cases under the bivariate non-stationarity assumption for $\{X_t, Y_t\}$. We investigate the properties of the new correlation coefficient in two special cases: namely, Spurious Correlation and Cointegration. Then, we simulate both the new and Pearson's correlation coefficients for these cases in Section 4.

Spurious correlation: in this case, X_t and Y_t are generated by the independent random walks:

$$\begin{aligned} X_t &= X_{t-1} + \varepsilon_t^x, \\ Y_t &= Y_{t-1} + \varepsilon_t^y, \end{aligned} \tag{16}$$

in which ε_t^x are $i.i.d.(0, \sigma_{\varepsilon^x}^2)$ and ε_t^y are $i.i.d.(0, \sigma_{\varepsilon^y}^2)$. In our simulations in Section 4, we draw ε_t^x and ε_t^y from independent $N(0, 1)$ populations. As T goes to infinity, the numerator of the new correlation coefficient in (6) goes to zero [6].

$$\alpha_{xy} = \mathbf{E}\left[(X_t - X_{t-1})\,(Y_t - Y_{t-1})\right] = \mathbf{E}\left[\left(\varepsilon_t^x\right)\left(\varepsilon_t^y\right)\right] \to 0. \tag{17}$$

Therefore, the new correlation coefficient also goes to zero.

$$\rho_O = \frac{\alpha_{xy}}{\alpha_x \alpha_y} \to 0 \quad \text{a.s.} \tag{18}$$

Cointegration: now we have two cointegrated non-stationary variables which are generated by,

$$\begin{aligned} X_t &= X_{t-1} + \varepsilon_t^x, \\ Y_t &= \alpha X_t + \varepsilon_t^y, \end{aligned} \tag{19}$$

in which ε_t^x are $i.i.d.(0, \sigma_{\varepsilon^x}^2)$ and ε_t^y are $i.i.d.(0, \sigma_{\varepsilon^y}^2)$. As T goes to infinity

$$\rho_O = \frac{\mathbf{E}\left[(X_t - X_{t-1})\,(Y_t - Y_{t-1})\right]}{\sqrt{\mathbf{E}\left[(X_t - X_{t-1})^2\right]\mathbf{E}\left[(Y_t - Y_{t-1})^2\right]}} \to \frac{1}{\sqrt{1 + \frac{2}{\alpha^2}\frac{\sigma_{\varepsilon^y}^2}{\sigma_{\varepsilon^x}^2}}} \tag{20}$$

since,

$$\begin{aligned} &\mathbf{E}\left[\left(\varepsilon_t^x\right)^2\right] = \sigma_{\varepsilon^x}^2 \quad \text{as } T \to \infty, \\ &\mathbf{E}\left[\left(\varepsilon_t^y\right)^2\right] = \mathbf{E}\left[\left(\varepsilon_{t-1}^y\right)^2\right] = \sigma_{\varepsilon^y}^2 \quad \text{as } T \to \infty. \end{aligned}$$

3.4. The properties in general context

Here, we introduce two propositions and show their proofs in general circumstances. The state of the independent variables and the essential properties of the new correlation coefficient enable us to make two propositions.

Proposition 1. *If X_t, Y_t are one lag cross-independent[4] and also if at least one of X_t and Y_t is not divergent,[5] then the new correlation coefficient $\rho_O = 0$.*

Proof. The numerator of ρ_O is α_{xy} and equals to:

$$\alpha_{xy} = \mathbf{E}\left[(X_t - X_{t-1})\,(Y_t - Y_{t-1})\right]. \tag{21}$$

Because of the linearity property of the expected value, Eq. (21) becomes as follows:

$$\alpha_{xy} = \mathbf{E}\left[X_t Y_t\right] - \mathbf{E}\left[X_t Y_{t-1}\right] - \mathbf{E}\left[X_{t-1} Y_t\right] + \mathbf{E}\left[X_{t-1} Y_{t-1}\right]. \tag{22}$$

Since X_t and Y_t are one lag cross-independent, the RHS of Eq. (22) turns into:

$$\alpha_{xy} = \mathbf{E}\left[X_t\right]\mathbf{E}\left[Y_t\right] - \mathbf{E}\left[X_t\right]\mathbf{E}\left[Y_{t-1}\right] - \mathbf{E}\left[X_{t-1}\right]\mathbf{E}\left[Y_t\right] + \mathbf{E}\left[X_{t-1}\right]\mathbf{E}\left[Y_{t-1}\right]. \tag{23}$$

[4] We define one lag cross independence between two variables X_t and Y_t, if pairs of $\{X_{t-m}, Y_{t-n}\}$ for $m, n = 0, 1$, are pairwise independent.

[5] Here, divergence of a variable Z_t is defined when $\mathbf{E}(Z_t) \neq \mathbf{E}(Z_{t-1})$.

By the non-divergence of X_t, $\mathbf{E}(X_t) = \mathbf{E}(X_{t-1})$. Eq. (23) can be rewritten as:

$$\alpha_{xy} = \mathbf{E}[X_t]\mathbf{E}[Y_t] - \mathbf{E}[X_t]\mathbf{E}[Y_{t-1}] - \mathbf{E}[X_t]\mathbf{E}[Y_t] + \mathbf{E}[X_t]\mathbf{E}[Y_{t-1}] \tag{24}$$

$$= 0. \tag{25}$$

Therefore, $\alpha_{xy} = 0$, implying $\rho_O = 0$. □

This proposition tells us that whenever there is at least one non-divergent variable, one lag cross-independence implies a zero new correlation coefficient. However, Pearson's correlation coefficient may not capture the cross-independence. As such, it may not be zero when at least one of the variables is not stationary. The condition of non-divergence is stronger than the stationarity condition. Not only stationary variables, but also some non-stationary variables such as random walks are non-divergent. Therefore, the new correlation coefficient is capable of observing the one lag cross-independence between the variables, even if the variables are non-stationary. Furthermore, the contra-positive of this proposition is also beneficial. If the new correlation coefficient is not zero, this refers to a violation of the one lag cross-independence condition or two divergent variables which are non-stationary.

Proposition 2. *If $\rho_O = 0$, then X_t, Y_t are not one lag cross-independent or at least one of X_t and Y_t is not divergent.*

Proof. Suppose for a contradiction that the imposed condition does not hold. So we have that X_t, Y_t are one lag cross-independent, and none of X_t and Y_t is not divergent, that is both of X_t and Y_t are divergent.

By the derivation of the previous proposition, the one lag cross-independence indicates that:

$$\alpha_{xy} = \mathbf{E}[X_t]\mathbf{E}[Y_t] - \mathbf{E}[X_t]\mathbf{E}[Y_{t-1}] - \mathbf{E}[X_{t-1}]\mathbf{E}[Y_t] + \mathbf{E}[X_{t-1}]\mathbf{E}[Y_{t-1}]. \tag{26}$$

Furthermore, divergent X_t and Y_t mean that:

$$\mathbf{E}(X_t) = \mathbf{E}(X_{t-1}) + \alpha \tag{27}$$

$$\mathbf{E}(Y_t) = \mathbf{E}(Y_{t-1}) + \beta \tag{28}$$

where α and β are any real number and not equal to zero.

Therefore, Eq. (26) may be rewritten as:

$$\alpha_{xy} = [\mathbf{E}[X_{t-1}] + \alpha][\mathbf{E}[Y_{t-1}] + \beta] - [\mathbf{E}[X_{t-1}] + \alpha]\mathbf{E}[Y_{t-1}] - \cdots \mathbf{E}[X_{t-1}][\mathbf{E}[Y_{t-1}] + \beta] + \mathbf{E}[X_{t-1}]\mathbf{E}[Y_{t-1}] \tag{29}$$

$$= \alpha\beta. \tag{30}$$

That is to say that the numerator of ρ_O is not zero ($\alpha_{xy} = \alpha\beta \neq 0$). Also, the denominator of ρ_O is positive, so ρ_O is not zero which contradicts our imposing statement that says that $\rho_O = 0$. □

Here, this proposition is similar to the converse of the previous proposition. We try to figure out what a zero new correlation coefficient implies. We find that if the new correlation coefficient is zero, then X_t and Y_t do not provide the condition of one lag cross-independence or at least one of the variables is not divergent. Therefore, a zero new correlation coefficient indicates that the one lag cross-independence condition is violated or that there is a minimum of one non-divergent variable.

4. Monte Carlo simulations

4.1. Stationary case

Here we report a Monte Carlo simulation to compare the new correlation coefficient and Pearson's correlation coefficient for one of the cases of the stationary variables which is stated in Section 3.2. In our Monte Carlo simulation setup, we form 5000 simulations and for each simulation 5000 data points are generated for both of the variables. We employ the case of two stationary variables which are both i.i.d. and independent of each other. The results of the simulation are given in Fig. 2.

As we mention in Section 3.2, both correlation coefficients go to zero because these variables are two i.i.d. stationary variables and independent of each other. When we compare the two correlation coefficients, we reach that the mean value of the new correlation coefficient is closer to zero than Pearson's correlation coefficient. So the new one is faster in reaching to zero compared to Pearson's one. Additionally, the standard deviation of the new correlation coefficient is less than the Pearson's correlation coefficient. As a result, the new correlation coefficient dominates Pearson's correlation coefficient in terms of capturing the independence among the variables.

4.2. Non-stationary cases

In this section, we perform Monte Carlo simulations to illustrate some of the properties of the new correlation coefficient compared to Pearson's correlation coefficient for various non-stationarity cases.

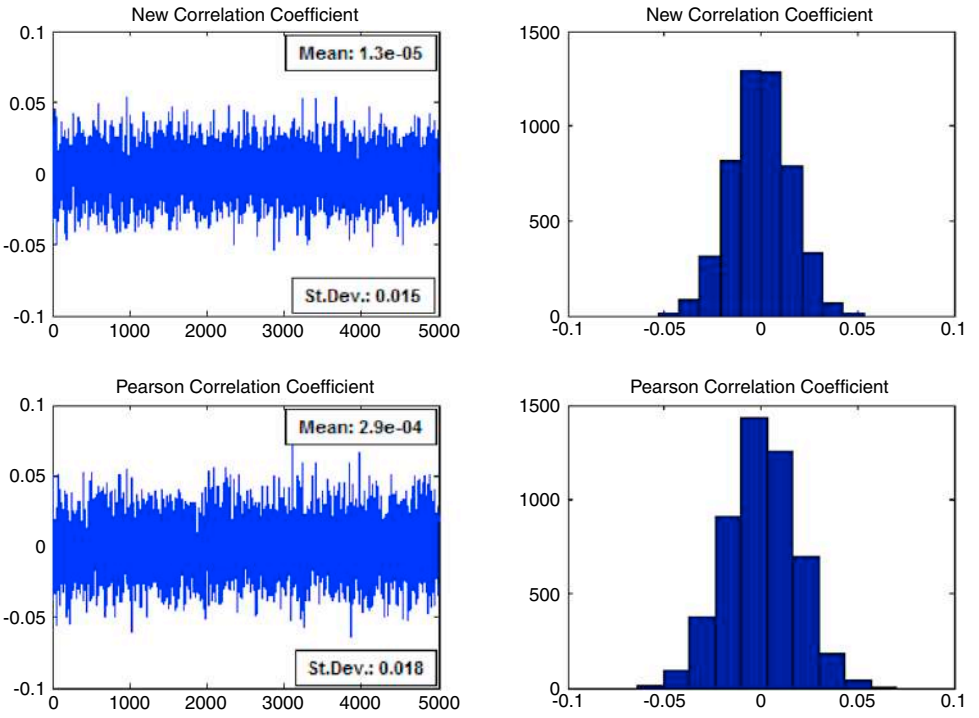

Fig. 2. Correlation coefficients in the case of independent stationary variables. Note: the plot in the top left and the histogram in the top right show how the new correlation coefficient is distributed. The plot and histogram in the bottom are for Pearson's correlation coefficient.

4.2.1. Two random walks (spuriously correlated)

Here, we take two nonstationary variables, such as two random walks:

$$X_t = X_{t-1} + \varepsilon_t^x, \quad \text{where } \varepsilon_t^x \sim N(0, 1)$$
$$Y_t = Y_{t-1} + \varepsilon_t^y, \quad \text{where } \varepsilon_t^y \sim N(0, 1)$$

which can be spuriously correlated or not cointegrated, also the distributions ε_t^x and ε_t^y are independent. Again, we form 5000 simulations and for each simulation 5000 data points are generated for both of the variables X_t and Y_t where X_0 and Y_0 are given as arbitrary numbers. The results of the simulations are shown in Fig. 3. The differences between the new and Pearson's correlation coefficients in the case of spurious correlation can be easily seen by looking at the plots and histograms of the coefficients.

In this case, it is expected that the correlation of the two nonstationary variables should be around zero because these variables are two independent random walks. Pearson's correlation coefficient on the average is zero, but has extremely large fluctuations (i.e., it could imply a very strong positive or negative correlation for some of the simulated data sets). This issue is called as spurious correlation in the literature [7]. However, the new correlation coefficient takes into account the nonstationarities of the variables, so it provides better results. The values of the simulations of the new correlation coefficient are more clustered around zero and have a shape more similar to normal distribution. Additionally, when we take two random walks with drift, we obtain similar results (not presented).

4.2.2. Two cointegrated variables

We take two nonstationary variables as:

$$X_t = X_{t-1} + \varepsilon_t^x, \quad \text{where } \varepsilon_t^x \sim N(0, 1)$$
$$Y_t = aX_t + \varepsilon_t^y, \quad \text{where } \varepsilon_t^y \sim N(0, 1)$$

in which $\{X_t\}$ and $\{Y_t\}$ are cointegrated. For $a = 0.1, 0.2, 0.3, \ldots, 1$, we use the same Monte Carlo simulation setup as in the previous case (in Section 4.2.1). Fig. 4 shows the box-plot of 5000 simulations for both the new and Pearson's correlation coefficients at each value of a. The value of the Pearson's correlation coefficient is significantly higher than the value of the new correlation coefficient for each value of a. These differences in Fig. 4 can be explained by the estimations of correlations and the parameters in these estimations. Actually, both of correlations capture the cointegration but give different values according to the parameters and a's. For example, the new correlation coefficient is calculated according to Eq. (20) above, and it is 0.58 when $a = 1$.

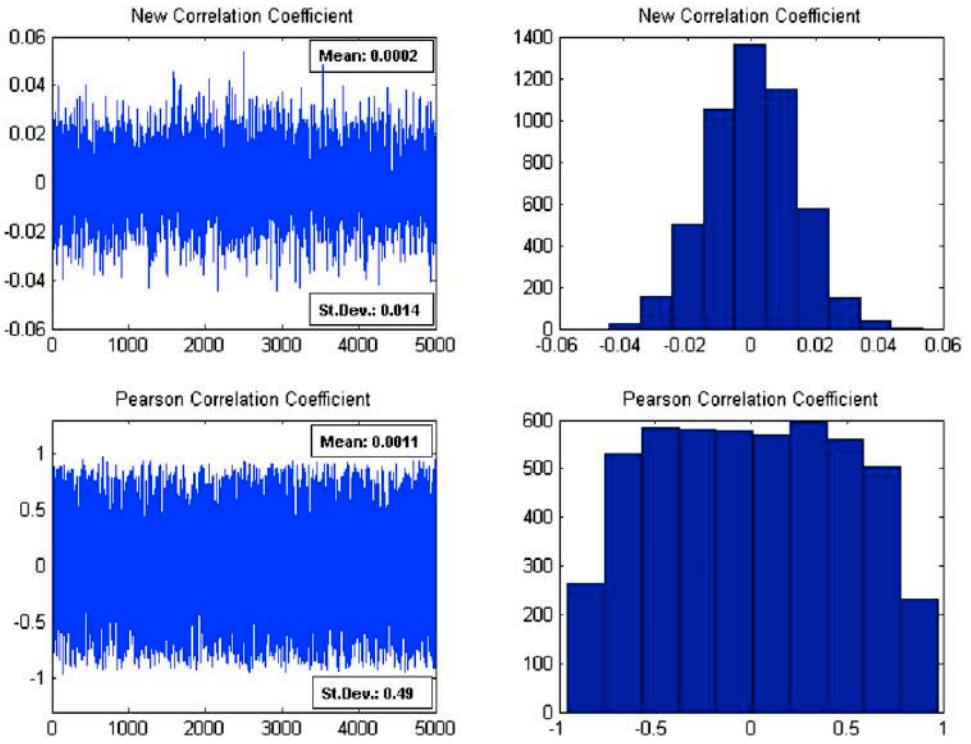

Fig. 3. Correlation coefficients in the case of spurious correlation. Note: the plot in the top left and the histogram in the top right show how the new correlation coefficient is distributed. The plot and histogram in the bottom are for Pearson's correlation coefficient.

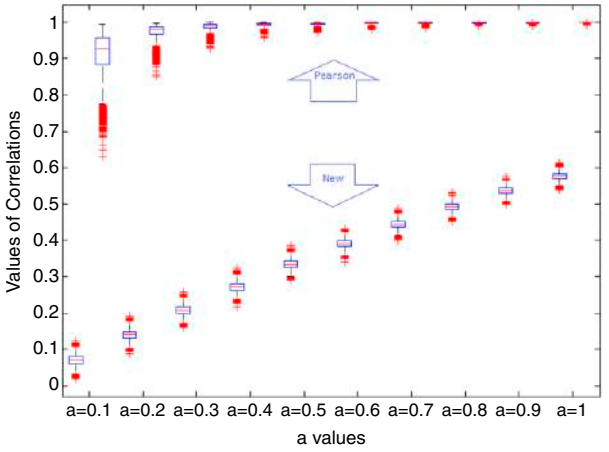

Fig. 4. Box-plots of correlation coefficients in the case of cointegration. Note: box-plots represent the new and Pearson's correlation coefficients for different a values.

4.3. Comparison of new correlation coefficient and DCCA

Correlations based on cross-covariance analysis between bivariate time series have become a trend in Finance and Economics over recent years. The main intuition behind the construction of these correlations is the separation of time series into overlapping boxes of length (scale). Then the correlation analysis is composed of aggregation of the information coming from separated time scales. One of the such correlations is called detrended cross-correlation DCCA (see Refs. [8,9]). For non-stationary time series, it is shown that the DCCA coefficient dominates Pearson's correlation coefficient (see Refs. [10–12]).[6] Here we also compare our results with the DCCA values for the two special cases of non-stationary variables.

[6] There are also alternative coefficients such as detrending moving-average cross-correlation analysis DMCA (see Refs. [13]), but DCCA fits for the analysis we use in this section.

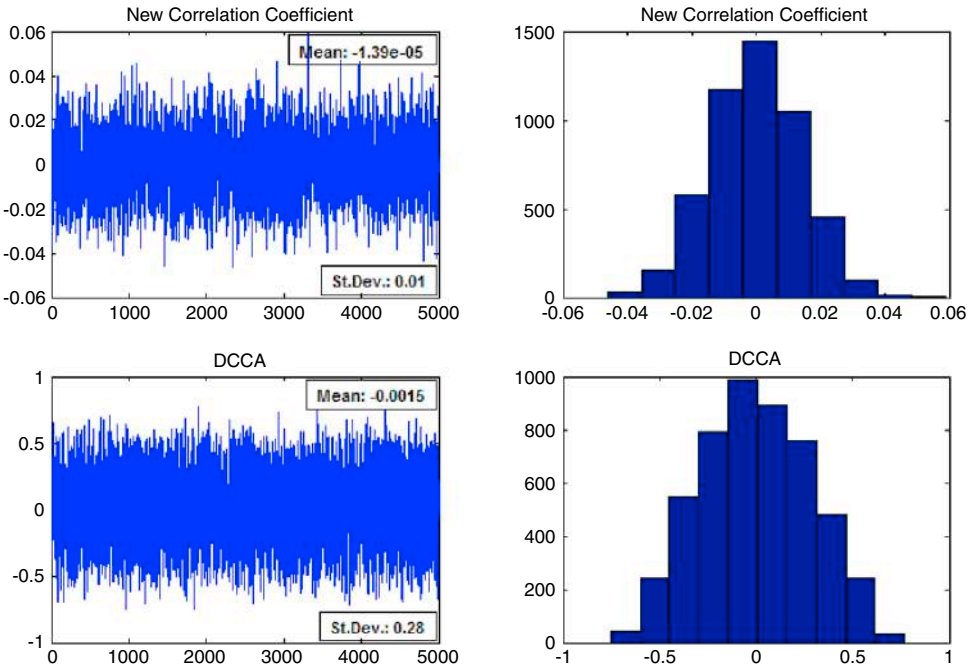

Fig. 5. Comparison of new and DCCA correlation coefficients in the case of spurious correlation. Note: the plot in the top left and the histogram in the top right show how the new correlation coefficient is distributed. The plot and histogram in the bottom are for Pearson's correlation coefficient.

First, we again take two independent nonstationary variables which are illustrated by two random walks. Similar to the simulation setup in the previous subsection, we perform 5000 simulations. For each simulation, 5000 data points are generated for both of the variables. The length of the overlapping boxes calculating the DCCA is chosen as the half of data points, which is 2500. The results of the simulations are shown in Fig. 5. When we compare the results for the new correlation with DCCA, we reach that both correlation coefficients capture zero correlation in the case of two independent random walks. Moreover, the standard deviation of the DCCA coefficient is found to be 0.28. The results in Ref. [12] also indicate that if there are two independent random walks, the standard deviation of the DCCA coefficient can take values from 0.08 to 0.40 with respect to changing length of boxes. However, the new correlation coefficient has standard deviation of 0.014 and five times less than the smallest standard deviation of DCCA coefficient. Therefore, the new correlation coefficient dominates the DCCA coefficient in the case of two independent random walks.

Additionally, we generate two cointegrated variables with the same setting described in the previous subsection. Both of the correlation coefficients again capture the relationship between the variables depending on the parameters selection. Kristoufek [12] illustrates that the standard deviation of the DCCA coefficient is in between 0.01 and 0.1 with respect to changing length of boxes. For the same case, the standard deviation of the new correlation coefficient is 0.01 which is the same as the smallest standard deviation of the DCCA coefficient. That is, the new correlation coefficient has almost similar performance with the DCCA coefficient in the case of cointegration. As a consequence, the new correlation coefficient captures the two special cases of nonstationarity mentioned above better than the DCCA correlation coefficient does in total.

5. Concluding remarks

Correlation in social sciences, especially in finance, has gained considerable attention recently. Several financial instruments and methods such as pair trading, credit risk applications, etc., focus on the correlation between various pairs of variables within data sets. One of the measures of inference for the correlation is Pearson's correlation coefficient, which requires the following assumptions: normality, linearity, randomness, stationarity and homoscedasticity. Unfortunately, these assumptions are rarely satisfied for real data. Pearson's correlation coefficient is not appropriate, if assumptions such as stationarity do not hold. The pitfalls of Pearson's correlation coefficient motivated us to introduce a new correlation coefficient which is measured by taking into account the lag difference of data points. We demonstrate that the new correlation coefficient has better performance in capturing the cross-independence of two variables over time. Additionally, we illustrated the key differences between Pearson's and the new correlation coefficients under a number of cases with Monte Carlo simulations. The most notable improvements of the new correlation coefficient can be seen in the situation of the two nonstationary variables which are spuriously correlated. That is, the new correlation coefficient captures independence of variables more easily than Pearson's coefficient does. Our simulation study suggests that the new correlation coefficient is more similar to normal distribution as compared to Pearson's correlation coefficient in the cases we consider. As a robustness

check, we also compare the new correlation coefficient with the DCCA coefficient. The comparison is simulated under two special cases for the non-stationary variables. According to the simulation results, the new correlation coefficient performs better than the DCCA coefficient in terms of capturing the independence and cointegration in non-stationary series.

Appendix. Correlation coefficients in the matrix form

An estimator of Pearson's correlation coefficient ρ_P is given by

$$\widehat{\rho}_P = \frac{\sum_{t=1}^{T} \left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)}{(T-1)s_x s_y}. \tag{31}$$

Numerator of $\widehat{\rho}_P$ in (21) can be written as

$$\widehat{\rho}_{P'} = (\mathbf{X}'\mathbf{A})(\mathbf{A}\mathbf{Y}) \tag{32}$$

where $\mathbf{X}' = (X_1, X_2, \ldots, X_T)$ and $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_T)$, and

$$\mathbf{A} = \begin{pmatrix} 1-\frac{1}{T} & -\frac{1}{T} & \cdot & \cdot & \cdot & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & \cdot & \cdot & \cdot & -\frac{1}{T} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{1}{T} \\ -\frac{1}{T} & \cdot & \cdot & \cdot & -\frac{1}{T} & 1-\frac{1}{T} \end{pmatrix}.$$

Here $\mathbf{A}$ is a $T \times T$ matrix, and $\mathbf{X}$ and $\mathbf{Y}$ are $T \times 1$ vectors, with $'$ standing for the transpose of a vector or matrix. Notice that

$$\mathbf{A}^2 = \begin{pmatrix} 1-\frac{1}{T} & -\frac{1}{T} & \cdot & \cdot & \cdot & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & \cdot & \cdot & \cdot & -\frac{1}{T} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{1}{T} \\ -\frac{1}{T} & \cdot & \cdot & \cdot & -\frac{1}{T} & 1-\frac{1}{T} \end{pmatrix}.$$

Hence, $\mathbf{A}$ is also an idempotent matrix. So, the estimate of Pearson's correlation coefficient in (21) can be written as

$$\widehat{\rho}_P = \frac{\mathbf{X}'\mathbf{A}\mathbf{Y}}{(T-1)\sqrt{\mathbf{X}'\mathbf{A}\mathbf{X}}\sqrt{\mathbf{Y}'\mathbf{A}\mathbf{Y}}}. \tag{33}$$

A similar formula can be constructed for our new correlation coefficient. Numerator in $\widehat{\rho}_o$ can be written as

$$\widehat{\rho}_{o'} = (T-1)A_{xy} = (\mathbf{C}\mathbf{X})'(\mathbf{C}\mathbf{Y}) = (\mathbf{X}'\mathbf{C}')(\mathbf{C}\mathbf{Y}) \tag{34}$$

where $\mathbf{X}' = (X_1, X_2, \ldots, X_T)$ and $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_T)$, and

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & -1 & 1 & 0 & \cdot & 0 \\ 0 & 0 & -1 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & -1 & 1 \end{pmatrix}.$$

Notice that $\mathbf{C}$ is a $(T-1) \times T$ matrix, and $\mathbf{X}$ and $\mathbf{Y}$ are $T \times 1$ vectors. We compute the product which is a $(T-1) \times (T-1)$ matrix, as

$$\mathbf{M} = \mathbf{C}' \cdot \mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ -1 & 2 & -1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & -1 & 2 & -1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & -1 & 2 & -1 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & -1 & 1 \end{pmatrix}.$$

Hence, we can reformulate the new correlation coefficient as

$$\widehat{\rho}_O = \frac{\mathbf{X}'\mathbf{M}\mathbf{Y}}{(T-1)\sqrt{\mathbf{X}'\mathbf{M}\mathbf{X}}\sqrt{\mathbf{Y}'\mathbf{M}\mathbf{Y}}}. \tag{35}$$

References

[1] R.R. Wilcox, J. Muska, Inferences about correlations when there is heteroscedasticity, Br. J. Math. Stat. Psychol. 54 (2001) 39–47.
[2] C. Alexander, Quantitative Methods in Finance, John Wiley & Sons, 2008.
[3] W. Enders, Applied Econometric Time Series, John Willey & Sons, NY, 2004.
[4] R.S. Tsay, Analysis of Financial Time Series, vol. 543, John Wiley & Sons, 2005.
[5] P.J. Brockwell, R.A. Davis, Time Series: Theory and Methods, Springer, 2009.
[6] P.C. Phillips, Understanding spurious regressions in econometrics, J. Econometrics 33 (1986) 311–340.
[7] R.F. Engle, C.W. Granger, Co-integration and error correction: representation, estimation, and testing, Econometrica (1987) 251–276.
[8] B. Podobnik, Z.Q. Jiang, W.X. Zhou, H.E. Stanley, Statistical tests for power-law cross-correlated processes, Phys. Rev. E 84 (2011) 066118.
[9] G. Zebende, M. Da Silva, A. Machado Filho, Dcca cross-correlation coefficient differentiation: theoretical and practical approaches, Physica A 392 (2013) 1756–1761.
[10] B. Podobnik, H.E. Stanley, Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series, Phys. Rev. Lett. 100 (2008) 084102.
[11] G. Zebende, Dcca cross-correlation coefficient: quantifying level of cross-correlation, Physica A 390 (2011) 614–618.
[12] L. Kristoufek, Measuring correlations between non-stationary series with dcca coefficient, Physica A 402 (2014) 291–298.
[13] L. Kristoufek, Detrending moving-average cross-correlation coefficient: measuring cross-correlations between non-stationary series, Physica A 406 (2014) 169–175.