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.

## 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.

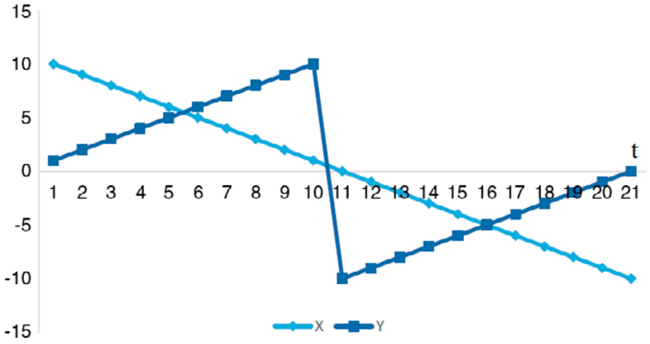

**Fig. 1.** An illustration of directionality detection problem. Note: $\begin{cases} X_t \text{ points (diamond) are generated as } X_t = 11 - t \text{ and} \\ Y_t \text{ points (square) are generated as } Y_t = t - 1 \text{ for } t = 1 \text{ to11 and } Y_t = t - 22 \text{ for } t = 12 \\ \quad \text{to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

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 $\rho_p$, 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 $t - s$:

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

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

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

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.