RESEARCH ARTICLE

Correlation and the time interval over which the variables are measured – A non-parametric approach

Edna Schechtman¹, Amit Shelef^{2*}

1 Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva, Israel, **2** Department of Industrial Management, Sapir Academic College, Sderot, Israel

* amitshe@mail.sapir.ac.il

Abstract

It is known that when one (or both) variable is multiplicative, the choice of differencing intervals (n) (for example, differencing interval of $n = 7$ means a weekly datum which is the product of seven daily data) affects the Pearson correlation coefficient (ρ) between variables (often asset returns) and that ρ converges to zero as n increases. This fact can cause the resulting correlation to be arbitrary, hence unreliable. We suggest using Spearman correlation (r) and prove that as n increases Spearman correlation tends to a limit which only depends on Pearson correlation based on the original data (i.e., the value for a single period). In addition, we show, via simulation, that the relative variability (CV) of the estimator of ρ increases with n and that r does not share this disadvantage. Therefore, we suggest using Spearman when one (or both) variable is multiplicative.

OPEN ACCESS

Citation: Schechtman E, Shelef A (2018) Correlation and the time interval over which the variables are measured – A non-parametric approach. PLoS ONE 13(11): e0206929. <https://doi.org/10.1371/journal.pone.0206929>

Editor: Vygtintas Gontis, Vilnius University, LITHUANIA

Received: June 4, 2018

Accepted: October 19, 2018

Published: November 8, 2018

Copyright: © 2018 Schechtman, Shelef. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1. Background

In the social science literature, the connection between two variables (X , Y) is often evaluated for a specific differencing interval (n). It has been quite common to arbitrarily pick differencing intervals (for example, weekly ($n = 7$), monthly ($n = 30$) or quarterly ($n = 90$) data when the data available are collected daily) and use sums or products in the analyses. For example, for a sample size of 210 daily observations, a weekly differencing interval of $n = 7$ for the multiplicative case will result in 30 weekly observations: $W'_{n=7,1}, W'_{n=7,2}, \dots, W'_{n=7,30}$, where $W'_{n=7,j} = \prod_{i=(j-1) \cdot 7 + 1}^{7j} X_i$ or, in general, $W'_{n,j} = \prod_{i=(j-1)n + 1}^{jn} X_i$ for $j = 1, 2, \dots, 30$. Multiplicative cases are used for rates of stocks, population size and more. However, this arbitrariness of choosing a differencing interval is dangerous as it may affect the correlation coefficient and the conclusions from the data. [1] writes: "Series can be used in correlation calculations as hourly, daily, weekly, quarterly or annual data. The resulting C/C/SV/V (Correlation, Covariance, Semi-Variance and Variance) coefficients will differ substantially for each time interval used. This is a major weakness—ideally, any accurate measure of co-movement should not be affected by the choice of time interval used for selecting variables".

It is easy to see that if two series are random walks, each consisting of the partial sums of a sequence of independent, identically distributed (i.i.d.) random variables (called additive-

additive model and denoted by aa), then Pearson correlation coefficient (ρ_n) between them will be independent of the differencing interval (n). However, there are cases where multiplicative models are called for. For example, growth rates of gross domestic products (GDP), industrial production, rates of stocks or population size.

For the multiplicative-multiplicative case (denoted by mm) [2] show that the limit of Pearson correlation coefficient (ρ_n) tends to 0 as the differencing interval increases. This holds under the assumptions that both variables are positive and have positive variances (except for the case when $Y = kX$ for some positive k in which the correlations are equal to 1 for all n) and that there is independence across time. In a recent book [3] notes that although $\lim_{n \rightarrow \infty} \rho_n = 0$, the speed of convergence of the (Pearson) correlation matrix to the diagonal matrix is not revealed in [2]. The third case, additive-multiplicative case (denoted by am or ma, depending on which variable is taken in its multiplicative form and which one is additive), was studied by [4]. They show that Pearson correlation coefficient (ρ_n) tends to 0 as the differencing interval increases.

[5] investigated a related problem—the effect of differencing intervals on simple regression coefficients. They illustrated that in any econometric study in which the variables have multiplicative rather than additive properties, the regression coefficients will have a mathematical bias, which is a function of the unit of time for which the empirical data are collected. [6] examined the effect of differencing intervals on multiple regression models with a combination of additive and multiplicative variables. They showed that the correlation and the partial regression coefficients approach zero as time interval increases.

The objective of this paper is twofold. The first objective is to investigate the effect of differencing interval on Spearman (r) correlation coefficient in the three models: additive-additive, multiplicative-multiplicative and additive-multiplicative. We adopt the assumptions used in [2] and [4], namely that the data are i.i.d. bivariate variables having second moments, and our focus is on the effect of the differencing interval. We prove that the n -period Spearman correlation converges to a limit as the differencing interval increases (while the n -period Pearson correlation converges to zero, as shown in [2] and [4]). The fact that the n -period Spearman correlation tends to a constant as n increases makes Spearman correlation a reliable measure from the theoretical point of view, because the user cannot “play” with the differencing interval to obtain a desired correlation. The second objective is to illustrate, via simulations, the median and the relative variability (CV) of the estimates for Pearson and Spearman correlations for a choice of underlying distributions.

The structure of the paper is the following: in Section 2 we present the main results of the paper, namely the limits of Spearman correlations as the differencing interval increases. In Section 3 we report on the simulation results with respect to the median and the CV of Pearson and Spearman correlations. Section 4 concludes.

2. The n -period Spearman correlation

Let $(X_1, Y_1), (X_2, Y_2), \dots$ be a sequence of i.i.d. random variables having a continuous bivariate distribution that meets the conditions of the central limit theorem (CLT). Denote $E(X) = \mu_X$, $V(X) = \sigma_X^2$, $E(Y) = \mu_Y$, $V(Y) = \sigma_Y^2$ and Pearson correlation coefficient ρ_{XY} . We start with the additive-additive (aa) case. Define the random variables $W_n = (W_{n,1}, W_{n,2}, \dots)$ and $V_n = (V_{n,1}, V_{n,2}, \dots)$, where

$$W_{n,j} = \sum_{i=(j-1)n+1}^{jn} X_i$$

and

$$V_{n,j} = \sum_{i=(j-1)n+1}^{jn} Y_i$$

for $j = 1, 2, \dots$

We denote their joint distribution function $F_{aa}^{(n)}(W_n, V_n)$. Using [7], there exists a copula $C_{aa}^{(n)}$ with uniform marginals such that

$$F_{aa}^{(n)}(W_n, V_n) = C_{aa}^{(n)}(H_a^{(n)}(W_n), G_a^{(n)}(V_n)),$$

where $H_a^{(n)}$ and $G_a^{(n)}$ are the marginal distribution functions of W_n and V_n , respectively.

Using the CLT, the distribution of the bivariate random variable

$$\left(\frac{W_n - n\mu_X}{\sqrt{n}\sigma_X}, \frac{V_n - n\mu_Y}{\sqrt{n}\sigma_Y} \right)$$

tends to a bivariate normal distribution with a mean vector (0,0), a variance vector (1,1) and a

correlation matrix $R = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{XY} & 1 \end{pmatrix}$ where ρ_{XY} is Pearson correlation coefficient between X_1

and Y_1 . Recall that for a bivariate normal random variable with a correlation matrix R , the Gaussian copula can be written as

$$\Phi_{\rho_{XY}}(\Phi^{-1}(u), \Phi^{-1}(v)),$$

where Φ_{ρ} is the bivariate normal distribution function with correlation ρ and Φ is the univariate normal cdf.

Therefore, the asymptotic dependence function of

$$\left(\frac{W_n - n\mu_X}{\sqrt{n}\sigma_X}, \frac{V_n - n\mu_Y}{\sqrt{n}\sigma_Y} \right)$$

is a Gaussian copula with correlation ρ_{XY} . Finally, copulas are invariant under strictly increasing transformations of the marginal distributions ([8], Theorem 2.4.3) so we get the following result:

Theorem 1: $C_{aa}^{(n)}$ tends to $\Phi_{\rho_{XY}}(\Phi^{-1}(u), \Phi^{-1}(v))$.

The relationship between Pearson and Spearman correlation coefficients was studied by [9] who showed that in the case of a bivariate normal distribution,

$$\rho = 2 \sin\left(\frac{\pi}{6} r\right)$$

or equivalently

$$r = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right)$$

where r is the Spearman correlation coefficient.

We now turn to the limit of the n -period rank correlation $r_{aa}^{(n)}$. By Theorem 1, the asymptotic copula of W_n and V_n is Gaussian. Therefore, using [9] result, the following theorem is immediate.

Theorem 2: $r_{aa}^{(n)}$ tends to $\frac{6}{\pi} \arcsin\left(\frac{\rho_{XY}}{2}\right)$.

We note in passing that it is easy to see that for the additive-additive case, Pearson correlation is not affected by the number of periods. That is, $\rho_{aa}^{(n)} = \rho_{XY}$ for all n .

We now extend the result to the two other cases: multiplicative-multiplicative (mm) and additive-multiplicative (am). We start with the multiplicative-multiplicative case as in [2]. Let $(X'_1, Y'_1), (X'_2, Y'_2), \dots$ be a sequence of i.i.d. random variables having a continuous bivariate distribution. Assume that X'_i and Y'_i are positive. Define the random variables $W'_n = (W'_{n,1}, W'_{n,2}, \dots)$ and $V'_n = (V'_{n,1}, V'_{n,2}, \dots)$, where

$$W'_{n,j} = \prod_{i=(j-1)n+1}^{jn} X_i$$

and

$$V'_{n,j} = \prod_{i=(j-1)n+1}^{jn} Y_i$$

for $j = 1, 2, \dots$

Following the additive-additive case, the asymptotic dependence of $(\ln(W'_n), \ln(V'_n))$ is the Gaussian copula with the dependence parameter being $\rho_{\ln(X')\ln(Y')}$. The following result for the n -period follows [9].

Theorem 3: $r_{mm}^{(n)}$ tends to $\frac{6}{\pi} \arcsin\left(\frac{\rho_{\ln(X')\ln(Y')}}{2}\right)$.

We note that because Spearman correlation is a rank correlation, $r(\ln(W'_n), \ln(V'_n)) = r(W'_n, V'_n)$, which completes the proof.

Recall that [2] show that as the number of periods n approaches infinity, $\rho_{mm}^{(n)}$ tends to zero (except for the case when Y is positively proportional to X , in which case the correlation is 1 for all n).

The third case, the additive-multiplicative case follows immediately.

Theorem 4: $r_{am}^{(n)}$ tends to $\frac{6}{\pi} \arcsin\left(\frac{\rho_{X\ln(Y')}}{2}\right)$.

The additive-multiplicative case was studied by [4]. They showed that $\rho_{am}^{(n)}$ tends to 0 as the differencing interval increases.

We note that the above theorems and proofs follow the same assumptions as in [2] and [4], namely that the data are i.i.d. bivariate variables having second moments. The focus here is on the effect of the differencing interval.

3. The effect of the differencing interval on the median and the relative variability (CV)

The second objective of this paper is to illustrate the effect of the differencing interval on the medians and the variabilities of the estimators of Pearson and Spearman correlation coefficients for a choice of differencing intervals via simulation. The additive-additive (aa) model will not be discussed in detail here because for this model the value of Pearson coefficient does not depend on the differencing interval. In addition, our simulation showed that in most cases its CV decreased with n . Therefore, Pearson correlation should be used for this model. The models under study are multiplicative-multiplicative (mm), following [2] and multiplicative-additive (ma), following [4].

In the simulation study we generated i.i.d. pairs (X, Y) , choosing a normal copula where the marginal distributions for X and Y were (Normal, Normal) (both distributions are symmetric),

Fig 1. Box plots for Pearson correlations for mm model with Lognormal and Normal distributions.

<https://doi.org/10.1371/journal.pone.0206929.g001>

(Lognormal, Normal) (one is symmetric and one is skewed), and (Lognormal, Lognormal) (both are skewed). The results of the three cases were similar, therefore we show the details of one case—the (Lognormal, Normal) case.

The simulation procedure is the following: In the first step we evaluated the “true” correlations. In order to do that we generated 1,000,000 pairs (X, Y) , choosing a normal copula where the marginal distributions for X and Y were Lognormal($0, 1.5^2$) (i.e., $\log(X)$ was Normal with mean 0 and variance 1.5^2) and Normal($1, 0.1^2$) (i.e., with mean 1 and variance 0.1^2), respectively. We added 0.1 to each observation in order to move away from zero (as required by [2]). The resulting Pearson and Spearman correlations were 0.32 and 0.58, respectively.

In the second step, we generated n pairs of (X, Y) from the distributions described in step 1 and computed the product and sum for each choice of n in order to obtain a single value of the bivariate variable (W'_n, V'_n) (or (W'_n, V_n)), according to mm (or ma) model, as detailed in section 2. The differencing intervals (n) were 1, 2, ..., 15, 20, 25, 30, 40, 50, 75 and 100. These values were chosen in order to cover the practical ranges: weekly, monthly or quarterly intervals for daily data, where $n = 7, 30$, and 90, respectively.

In the third step, the second step was repeated $R = 1000$ times in order to obtain 1000 replications of the bivariate variable (W'_n, V'_n) (or (W'_n, V_n)), namely $\{(W'_{n,1}, V'_{n,1}), \dots, (W'_{n,R}, V'_{n,R})\}$ (or $\{(W'_{n,1}, V_{n,1}), \dots, (W'_{n,R}, V_{n,R})\}$). Then, estimates of ρ and r were calculated based on R pairs of observations.

Fig 2. Box plots for Spearman correlations for mm model with Lognormal and Normal distributions.

<https://doi.org/10.1371/journal.pone.0206929.g002>

In the fourth step, the previous steps were repeated $M = 2000$ times to result in M estimates of ρ and r for each choice of n .

The criteria for comparison, for each n are the median, reported because the distributions of the estimates of Pearson correlation are skewed (as shown below), and the coefficient of variation (CV), based on M values for each correlation coefficient. A desirable property of a correlation coefficient is stability with respect to the differencing interval. That is, the median for n -period differencing should be similar to the median of the original data (i.e., the values for a single period) for all n , and the CV should be as small as possible.

Figs 1 to 4 illustrate the comparison between the two correlation coefficients. The box plots in Figs 1 and 2 present the results for Pearson and Spearman, respectively, for the mm model. As can be seen, the median of the estimates of ρ decreases as n gets larger, as expected. Note that the median for $n = 100$ is still 0.11. However, large n values are needed in order to approach 0 (the value for $n = 1500$ is 0.007, not shown in the figure). As mentioned above, similar results were obtained for all cases under study. In addition, the estimates of ρ vary from negative values to 0.8, and the distribution becomes skewed as n gets larger.

The CV (in percent) is shown in Figs 3 and 4 for the mm model and the ma model, respectively. As can be seen, the CV of the estimates of ρ increases as n gets larger (up to about 80% for $n = 100$ in Fig 3, for mm model).

We now turn to Spearman r . As opposed to Pearson coefficient, which converges to 0, Spearman r converges to a known constant, given in Theorem 3. Based on 1,000,000 pairs of

Fig 3. CV (in percent) of the correlation coefficients for mm model vs. n .

<https://doi.org/10.1371/journal.pone.0206929.g003>

observations sampled from the bivariate distribution detailed above, this value is equal to 0.57 for this case. As can be seen from Fig 2, r converges to this value and its distribution is quite symmetric. In addition, its CV is relatively low (Fig 3) and is less than 8% for all cases under study. We note that products of random variables can result in very extreme observations. This phenomenon may have a huge impact on Pearson correlation, but only a moderate one on Spearman correlation. This fact explains the a-priori advantage of using the Spearman correlation.

The simulation study that was used for the case of homoscedasticity was repeated for the case of heteroscedasticity and the results are given in the S1 Appendix. Results in the case of heteroscedasticity were similar to the case of homoscedasticity with slight differences in the rate of convergence and in the CV, as detailed in the S1 Appendix.

Conclusions

It is well documented that the differencing interval has an effect on Pearson correlation coefficient between two variables in multiplicative-multiplicative and multiplicative-additive models when $(X_1, Y_1), (X_2, Y_2), \dots$ is a sequence of i.i.d. random variables (for the multiplicative-multiplicative model they have to be non-negative). The coefficient tends to 0 as the differencing interval increases. A natural competitor, suggested in this study, is Spearman correlation

Fig 4. CV (in percent) of the correlation coefficients for ma model vs. n .

<https://doi.org/10.1371/journal.pone.0206929.g004>

coefficient, because the distribution of the multiplicative variable is often skewed, making Pearson coefficient less suitable.

We prove that Spearman correlation converges to a limit as the differencing interval increases in the three models: additive-additive, additive-multiplicative and multiplicative-multiplicative. The limits are functions of the Pearson correlation between the original (single period) variables. In addition, we show (via simulation) that for the multiplicative-additive and multiplicative-multiplicative models the distribution of the estimate of Pearson correlation becomes skewed and its CV increases as n gets larger, while Spearman correlation does not share this disadvantage. In the additive-additive model Pearson correlation is constant over all differencing intervals and its CV decreases as n increases. Therefore, Pearson correlation is preferred for this model. The simulations were repeated for the heteroscedastic case. Results show similar patterns with slight differences.

Therefore, we suggest the use of Spearman correlation for both the multiplicative-multiplicative and the additive-multiplicative models.

Supporting information

S1 Appendix. The effect of the differencing interval on the median and the relative variability (CV) for the case of heteroscedasticity. (DOCX)

Author Contributions

Conceptualization: Edna Schechtman, Amit Shelef.

Formal analysis: Edna Schechtman, Amit Shelef.

Investigation: Edna Schechtman, Amit Shelef.

Methodology: Edna Schechtman, Amit Shelef.

Software: Edna Schechtman, Amit Shelef.

Supervision: Edna Schechtman, Amit Shelef.

Validation: Edna Schechtman, Amit Shelef.

Writing – original draft: Edna Schechtman, Amit Shelef.

Writing – review & editing: Edna Schechtman, Amit Shelef.

References

1. Nwogugu M. On regret theory, and framing anomalies in the Net-Present-Value and the Mean-Variance models. 2014. Available from: <https://ssrn.com/abstract=2375022> or <http://dx.doi.org/10.2139/ssrn.2375022>.
2. Levy H, Schwarz G. Correlation and the time interval over which the variables are measured. *Journal of econometrics*. 1997; 76(1):341–50.
3. Levy H. Stochastic dominance: Investment decision making under uncertainty. 3rd ed: Springer; 2016.
4. Levy H, Guttman I, Tkatch I. Regression, correlation, and the time interval: Additive-multiplicative framework. *Management Science*. 2001; 47(8):1150–9.
5. Levhari D, Levy H. The capital asset pricing model and the investment horizon. *The Review of Economics and Statistics*. 1977:92–104.
6. Jea R, Lin J, Su C. Correlation and the time interval in multiple regression models. *European journal of operational research*. 2005; 162(2):433–41.
7. Sklar M. Fonctions de repartition an dimensions et leurs marges. *Publ Inst Statist Univ Paris*. 1959; 8:229–31.
8. Nelsen RB. An introduction to copulas. 2nd ed. New York: Springer; 2006.
9. Kruskal WH. Ordinal measures of association. *Journal of the American Statistical Association*. 1958; 53(284):814–61.