연세대학교 상경대학

경제연구소

Economic Research Institute
Yonsei Universit

서울시 서대문구 연세로 50
50 Yonsei-ro, Seodaemun-gu, Seoul, Korea
TEL: (+82-2) 2123-4065 FAX: (+82-2) 364-9149
E-mail: yeri4065@yonsei.ac.kr
http://yeri.yonsei.ac.kr/new

The Instability of the Pearson Correlation Coefficient in the Presence of Coincidental Outliers

Yunmi Kim
University of Seoul
Tae-Hwan Kim
Yonsei University
Tolga Ergün
State Street Corporation

Decemver 2014
2014RWP-74

Economic Research Institute
Yonsei University

The Instability of the Pearson Correlation Coefficient in the Presence of Coincidental Outliers

Yunmi Kim[a], Tae-Hwan Kim[b,*] and Tolga Ergün[c]

[a]Department of Economics, University of Seoul, Korea
[b]School of Economics, Yonsei University, Korea
[c]State Street Corporation, USA

October 2014

Abstract: It is well known that any statistic based on sample averages can be sensitive to outliers. Some examples are the conventional moments-based statistics such as the sample mean, the sample variance, or the sample covariance of a set of observations on two variables. Given that sample correlation is defined as sample covariance divided by the product of sample standard deviations, one might suspect that the impact of outliers on the correlation coefficient may be neither present nor noticeable because of a 'dampening effect' i.e., the effects of outliers on both the numerator and the denominator of the correlation coefficient can cancel each other. In this paper, we formally investigate this issue. Contrary to such an expectation, we show analytically and by simulations that the distortion caused by outliers in the behavior of the correlation coefficient can be fairly large in some cases, especially when outliers are present in both variables at the same time. These outliers are called 'coincidental outliers.' We consider some robust alternative measures and compare their performance in the presence of such coincidental outliers.

Keywords: Correlation; robust statistic; outliers

JEL classifications: C13, C18

*Corresponding author. Yonsei University, School of Economics,134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Korea; Tel.: +82-2-2123-5461; fax: +82-2-2123-8638; e-mail address: tae-hwan.kim@yonsei.ac.kr.
We are grateful to an anonymous referee and the editor Ramazan Gencay for their invaluable comments which substantially improved the paper.

1 Introduction

The sample correlation coefficient is probably the most frequently-used statistic for measuring the linear co-movement between two variables. It has been documented (e.g., see Stigler, 1989) that the essential idea of correlation or 'co-relation' was conceived by Francis Galton and was formally developed by Karl Pearson, which explains why it is sometimes called the 'Pearson correlation coefficient.' Although the correlation coefficient does not measure the causal relationship between two variables, it plays an important role in many scientific areas. For example, understanding how financial assets are moving together which is measured by the correlation coefficient is crucial in lowering portfolio risk through diversification.

Based on the main idea put forward by Kim and White (2004), Bonato (2010), Ergun (2011) and White et al. (2011), an intuitively appealing and easily computable robust measure of covariance has been proposed by Huo et al. (2012). They demonstrated that the conventional measure of covariance is heavily influenced by outliers. Given that sample correlation is defined as sample covariance divided by the product of two sample standard deviations, one might suspect that the impact of outliers on the correlation coefficient may not be either present or noticeable because of a 'dampening effect,' i.e., the effects of outliers on both the numerator and the denominator of the correlation coefficient can cancel each other.

In this paper, we formally investigate this issue. We first derive the analytical expression of the distortion caused by outliers. Then we attempt to gauge the size of such a distortion in many different situations using Monte Carlo simulations. As expected, there is a 'dampening effect' due to the standardization in many cases under consideration. However, there also exists a surprising twist in other cases, in particular when outliers are present in both variables at the same time. These outliers are called 'coincidental outliers.' In such cases, the discovered distortion is fairly large. We consider some robust alternative measures and compare their performance in the presence of such coincidental outliers.

2 The Effect of Outliers on the Conventional Measure of Correlation

We consider two stochastic processes $\{x_t\}_{t=1,\cdots,T}$ and $\{y_t\}_{t=1,\cdots,T}$ where x_t are assumed to be independent and identically distributed (IID) with the cumulative distribution function (CDF) F_x and y_t are also assumed to be IID with the CDF F_y. The conventional measure of correlation (denoted by C), called the 'Pearson correlation coefficient,' is given by:

$$C = \frac{E[(x_t - \mu_x)(y_t - \mu_y)]}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}}, \tag{1}$$

where $\mu_x = E(x_t)$, $\mu_y = E(y_t)$, $\sigma_x^2 = E[(x_t - \mu_x)^2]$, $\sigma_y^2 = E[(y_t - \mu_y)^2]$, and the expectation E is taken with respect to the joint CDF of x_t and y_t. The conventional measure C is, of course, a population parameter and thus must be estimated. Its usual estimation is achieved by replacing the population expectation E with its corresponding sample mean:

$$\widehat{C} = \frac{\frac{1}{T}\sum_{t=1}^{T}[(x_t - \widehat{\mu}_x)(y_t - \widehat{\mu}_y)]}{\sqrt{\widehat{\sigma}_x^2}\sqrt{\widehat{\sigma}_y^2}}, \tag{2}$$

where $\widehat{\mu}_x = \frac{1}{T}\sum_{t=1}^{T} x_t$, $\widehat{\mu}_y = \frac{1}{T}\sum_{t=1}^{T} y_t$, $\widehat{\sigma}_x^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \widehat{\mu}_x)^2$, and $\widehat{\sigma}_y^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{\mu}_y)^2$.

The above sample correlation $\widehat{C}$ is based on sample averages; thus, it may be influenced by any outliers in either x_t or y_t as explained in Kim and White (2004) and Huo et al. (2012). To determine the influence of outliers on the conventional measure, we assume without loss of generality that a single outlier (denoted as m_x) occurs at time $[\tau T]$ with $\tau \in (0,1)$ in x_t, and y_t also has an outlier (m_y) at time $[sT]$ with $s \in (0,1)$.[1] The case without outliers corresponds to when all of the observations on x_t and y_t are drawn from the joint distribution of x_t and y_t. On the other hand, the case with outliers are obtained by replacing $x_{[\tau T]}$ and $y_{[sT]}$ with $x_{[\tau T]} + m_x$ and $y_{[sT]} + m_y$, respectively. With these outliers, the sample correlation becomes:

$$\widehat{C} = \frac{\widehat{Cov}(x,y) + A(m_x, m_y)}{\left(\widehat{\sigma}_x^2 \widehat{\sigma}_y^2\right)^{1/2} \left(1 + \frac{B(m_x,m_y)}{\widehat{\sigma}_x^2 \widehat{\sigma}_y^2}\right)^{1/2}}, \tag{3}$$

where $\widehat{Cov}(x,y) = \frac{1}{T}\sum_{t=1}^{T}[(x_t - \widehat{\mu}_x)(y_t - \widehat{\mu}_y)]$ is the sample covariance computed without outliers. Similarly, $\widehat{\sigma}_x^2$ and $\widehat{\sigma}_y^2$ are the sample variances computed without outliers. Hence, the ratio $\widehat{C}_0 = \frac{\widehat{Cov}(x,y)}{\left(\widehat{\sigma}_x^2 \widehat{\sigma}_y^2\right)^{1/2}}$ is the sample correlation coefficient computed without the two outliers m_x and m_y. The distance between $\widehat{C}$ and $\widehat{C}_0$ is a measure of the distortion caused by the outliers. The other two terms appearing in (3) are defined as follows:

$$A(m_x, m_y) = \frac{m_y}{T}(x_{[sT]} - \widehat{\mu}_x) + \frac{m_x}{T}(y_{[\tau T]} - \widehat{\mu}_y) - \frac{m_x m_y}{T^2} + \frac{m_x m_y}{T}\mathbf{1}_{[[\tau T]=[sT]]}, \tag{4}$$

and

$$\begin{aligned} B(m_x, m_y) \quad = \quad & \frac{T-1}{T^2} m_x^2 \widehat{\sigma}_y^2 + 2\frac{1}{T} m_x (x_{[\tau T]} - \widehat{\mu}_x)\widehat{\sigma}_y^2 + \frac{T-1}{T^2} m_y^2 \widehat{\sigma}_x^2 + \frac{(T-1)^2}{T^4} m_x^2 m_y^2 \\ & + 2\frac{T-1}{T^3} m_x m_y^2 (x_{[\tau T]} - \widehat{\mu}_x) + 2\frac{1}{T} m_y (y_{[sT]} - \widehat{\mu}_y)\widehat{\sigma}_x^2 \\ & + 2\frac{T-1}{T^3} m_x^2 m_y (y_{[sT]} - \widehat{\mu}_y) + 4\frac{1}{T^2} m_x m_y (x_{[\tau T]} - \widehat{\mu}_x)(y_{[sT]} - \widehat{\mu}_y). \end{aligned} \tag{5}$$

The mathematical derivation of the result in (3) is straightforwardly obtained using the main results in Huo et al. (2012), hence the proof is omitted. We note that the sample correlation in (3) influenced by outliers can be expressed as:

$$\widehat{C} = \frac{\widehat{Cov}(x,y) + A(m_x, m_y)}{\left(\widehat{\sigma}_x^2 \widehat{\sigma}_y^2\right)^{1/2}} f(m_x, m_y), \tag{6}$$

where $f(m_x, m_y) = \left(1 + \frac{B(m_x,m_y)}{\widehat{\sigma}_x^2 \widehat{\sigma}_y^2}\right)^{-1/2}$. Then the first-order Taylor-series approximation of $f(m_x, m_y)$ around $(0,0)$ provides the following:

$$f(m_x, m_y) \approx 1 + f_{m_x}(0,0)m_x + f_{m_y}(0,0)m_y. \tag{7}$$

By substituting the Taylor approximation result in (7) into (6), we obtain:

$$\widehat{C} \approx \widehat{C}_0 + \widehat{D},$$

where

$$\widehat{D} = \frac{A(m_x, m_y)}{\left(\widehat{\sigma}_x^2 \widehat{\sigma}_y^2\right)^{1/2}} + \frac{\widehat{Cov}(x,y) + A(m_x, m_y)}{\left(\widehat{\sigma}_x^2 \widehat{\sigma}_y^2\right)^{1/2}} \left(f_{m_x}(0,0)m_x + f_{m_y}(0,0)m_y\right).$$

Since the first term $\widehat{C}_0$ is the sample correlation coefficient computed without outliers, the second term $\widehat{D}$ is the distortion caused by the presence of outliers. It measures how much the corrupted correlation coefficient $\widehat{C}$ deviates from the correct correlation coefficient $\widehat{C}_0$. The distortion term is a complicated function of the magnitudes of the two outliers (m_x and m_y) and it is difficult to derive an intuitive interpretation of the distortion term. Hence, we evaluate this distortion term by Monte Carlo simulations in Section 4 below.

[1] Note that the function $[a]$ is the usual integer function taking the integer part of the real number a.

3 Alternative Robust Measures

The most frequently-used robust measure for correlation in the statistics literature is the Spearman rank correlation, which is defined as:

$$\widehat{S} = 1 - \frac{6 \sum\limits_{t=1}^{T} d_t^2}{T(T^2 - 1)}, \tag{8}$$

where d_t is the difference in the ranks of x_t and y_t.[2] This expression implicitly assumes that there are no tied ranks.

Following the idea behind Kim and While (2004), Bonato (2010), Ergun (2011) and White et al. (2011), we consider another possible robust measure by replacing the population expectation E with a particular quantile *median*, which will be denoted as M. The expected value (E) and the median (M) of a random variable can equally measure the center of the distribution of the random variable, but the latter is more robust than the former in the presence of outliers. Therefore, the resulting robust measure of correlation (denoted by C_R) is given by:

$$C_R = \frac{M[(x_t - \kappa_x)(y_t - \kappa_y)]}{\sqrt{M[(x_t - \kappa_x)^2]}\sqrt{M[(y_t - \kappa_y)^2]}}, \tag{9}$$

where M is taken with respect to the joint CDF of x_t and y_t, and κ_x, κ_y are the population medians of x_t, y_t; i.e., $\kappa_x = M(x_t)$, $\kappa_y = M(y_t)$. A natural estimator for C_R (denoted as $\widehat{C}_R$) is given by the following:

$$\widehat{C}_R = \frac{\widehat{M}[(x_t - \widehat{\kappa}_x)(y_t - \widehat{\kappa}_y)]}{\sqrt{\widehat{M}[(x_t - \widehat{\kappa}_x)^2]}\sqrt{\widehat{M}[(y_t - \widehat{\kappa}_y)^2]}}, \tag{10}$$

where $\widehat{M}$ is the sample median operator (i.e., $\widehat{M}$ takes the middle value from a sample), and $\widehat{\kappa}_x, \widehat{\kappa}_y$ are the sample medians of x_t, y_t; i.e., $\widehat{\kappa}_x = \widehat{M}(x_t)$, $\widehat{\kappa}_y = \widehat{M}(y_t)$.[3] We note that $\widehat{C}_R$ corresponds to $\widehat{C}$ in (2) obtained by using the sample median in place of the sample mean.

The conventional measure of correlation C depends heavily on some moment conditions for the underlying distributions of x_t and y_t. For example, if the distribution of the product $(x_t - \mu_x)(y_t - \mu_y)$ does not have the first moment (as in the Cauchy distribution), then the conventional measure does not even exist. Of course, the same reasoning applies to either $(x_t - \mu_x)^2$ or $(y_t - \mu_y)^2$. However, the median-based robust measure in (10) is not susceptible to such moment conditions; in fact, the median-based measure is well-defined for any underlying distributions of x_t and y_t.

When comparing the above two robust measures, we note that if there are many tied ranks, the Spearman correlation can provide a poor approximation to what it intends to measure. On the other hand, the median-based correlation estimator in (10) is not vulnerable to such a criticism. We also note

[2] As shown in Schweizer and Wolff (1981), the population version (denoted as S) of the Spearman rank correlation in (8) is given by

$$S = 12 \int_{\infty}^{\infty}\int_{\infty}^{\infty} [F_{xy}(u,v) - F_x(u)F_y(v)]\, dF_x(u) dF_y(v),$$

where F_{xy} is the joint distribution of x_t and y_t.

[3] The conventional and robust correlation estimators ($\widehat{C}$, $\widehat{C}_R$) are based on means and medians, respectively. It is obvious that they do not estimate exactly the same population value although both mean and median are generally regarded as measuring the 'center' of the underlying distribution. A special case where these two measures estimate the same value is that x_t and y_t are independent. It turns out that $\widehat{C}_R$ tends to be smaller than $\widehat{C}$ in more general cases (i.e., when x_t and y_t are correlated). A detailed discussion about this point will be provided later in Section 4.

that the formula for the Spearman correlation is not readily interpretable whereas that for the median-based estimator is fairly intuitive. However, the relative computational merits of the two estimators are yet to be investigated, which will be carried out in the following section.

4 Monte Carlo Simulations

In this section, we conduct Monte Carlo simulations to investigate the behavior of the conventional measure ($\widehat{C}$), the median-based measure ($\widehat{C}_R$) and the Spearman measure ($\widehat{S}$) in the presence of outliers. The Monte Carlo simulations closely follow the simulation design of Kim and White (2004), Bonoto (2010) and Huo et al. (2012). We consider the standard normal distribution $N(0,1)$ and the student t-distributions with 5 and 1 degrees of freedom [T-5, T-1] to represent thin, moderately heavy and extremely heavy tailed symmetric distributions, respectively. For an asymmetric distribution, we use a lognormal distribution with $\mu = 1, \sigma = 0.4$ (centered at zero by subtracting the mean value) denoted by $LN(1, 0.4)$. It is known that T-1, also called the Cauchy distribution, has no first or higher moments. The inclusion of T-1 has been motivated by recent studies. For example, Bonato (2010) has shown that some time-series data can be regarded as coming from a distribution with infinite moments, and Ergun (2011) has provided some empirical evidence that the first few moments of S&P500 do not exist for some time periods.

We generate two processes $\{x_t\}_{t=1,\cdots,T}$ and $\{y_t\}_{t=1,\cdots,T}$ independently from the above four distributions with sample sizes $T = 50, 100, 300, 500$ and calculate the conventional and robust measures. We repeat the experiment 1,000 times to generate 1,000 estimates of both the conventional and robust measures. Because we draw these two processes $\{x_t\}_{t=1,\cdots,T}$ and $\{y_t\}_{t=1,\cdots,T}$ independently, the true value of correlation is obviously zero. If there is no outlier, all three measures should be well centered at zero, the true value.

Figures 1-4 show the results for the conventional measure ($\widehat{C}$). Each figure is constructed by four windows representing four distributions, and each window shows four boxplots for four different sample sizes. The vertical axis in each boxplot indicates the sample size. Each box-plot represents the lower (25%), median (50%), and upper (75%) quartiles. The size of both whiskers extending from the box is set to be the same as the size of each box (i.e., the inter-quartile range). The numbers at the end of each whisker indicate how many observations lie beyond the end of that whisker.

Figure 1 shows the performance of $\widehat{C}$ when there is no outlier. Under all four distributions ($N(0,1)$, T-5, T-1, $LN(1, 0.4)$), the conventional measure shows good performance and converges to zero (the true value) reasonably quickly. Interestingly, the conventional measure performs well even under T-1, which does not have any moment. According to Huo et al. (2012), the performance of the conventional measure of covariance is severely distorted under T-1. It seems that there might be some 'dampening effect' by the denominator of the correlation coefficient; i.e., the sample variances of x_t and y_t must be fairly large under T-1.

Figure 2 shows the case in which there is a single outlier only in x_t. Following Kim and White (2004), we used 48.62 times of the 25^{th} percentile of sample $\{x_t\}_{t=1,\cdots,T}$ as an outlier and we replaced the observation $x_{[0.3*T]}$ with this value in each sample. Kim and White (2004) have determined the size of outliers by considering the 1987 stock market crash. The worst daily loss was -20.41% during the crisis. The outliers in the simulations are constructed in such a way that they are comparable to such a massive market crash.[4] As shown in Figure 2, the single outlier does not seem to have any effect on the conventional measure.

We also consider the case in which there are outliers in both x_t and y_t at different times. An outlier for y_t is generated in the same way as that for x_t. That is, the value of the outlier for y_t is 48.62 times of the 25^{th} percentile of sample $\{y_t\}_{t=1,\cdots,T}$, and the observation $y_{[0.6*T]}$ is replaced by the outlier such that the two outliers occur at different times. The simulation results are shown in Figure 3. Interestingly, there exists some impact in this case, not in the distorting way, but in the way of 'improving' the performance

[4] See Kim and White (2004) for a more detailed explanation of the size of outliers.

of the measure because the sampling distribution of $\widehat{C}$ becomes much more concentrated around the true value, 0. This 'improvement' rather than 'distortion' must have been caused by the 'dampening effect.'

Figure 4 shows the final case; outliers occur at the same time which are called 'coincidental outliers.' The figure clearly visualizes the distortion caused by such outliers. The discovered distortion is spectacular. For the three distributions ($N(0,1)$, T-5, $LN(1,0.4)$), the center of the sampling distribution is very close to the maximum value, 1, and it is heavily tightened and concentrated around the center when $T=50$. The sampling distribution moves toward zero as T increases, but it moves fairly slowly so that the center is still either around or above 0.6 even for $T=500$. Also, the entire sampling distribution of the conventional measure far exceeds zero, indicating that researchers will incorrectly accept a spurious correlation 100% of the time. The discovered results shown in Figure 4 therefore may indicate that some uncorrelated or weakly correlated stocks may appear to be strongly correlated during a large financial crisis. Such an observation is consistent with the important asymmetric correlation literature[5] which has put forward a stylized fact that correlations between equities or stocks tend to increase during stressful market situations.[6]

We must also point out a potential problem with our simulation design, especially with coincidental outliers. The problem is that coincidental outliers are injected deterministically into the two independently generated series x_t and y_t. In reality, coincidental outliers are hardly viewed as deterministic because if the two series are independent, then the chance of a large shock hitting both on exactly the same time may be too small to be of any practical concern except a very rare event such as the 9/11 attack. Other events hitting the stock market like the 1987 stock market crash or the subprime mortgage crisis are better described as endogenously or stochastically occurring. However, implementing this more realistic design can be complicated and is beyond the scope of the paper. Hence, we must emphasize that the current simulation setup should be regarded as a convenient short-cut to examine the influence of simultaneous crashes that sometimes occur stochastically due to the strong dependence at the extreme tails of the underlying distributions.[7]

We now consider the behavior of the median-based robust measure $\widehat{C}_R$ and the Spearman measure $\widehat{S}$ in the presence of coincidental outliers.[8] Figures 5-6 display the behavior of the robust measures $\widehat{C}_R$ and $\widehat{S}$. The sampling distribution of $\widehat{C}_R$ is highly concentrated around zero as shown in Figure 5. Hence, the median-based measure $\widehat{C}_R$ does not lead to an incorrect conclusion based on a spurious correlation. The behavior of the Spearman rank correlation coefficient ($\widehat{S}$) is shown in Figure 6, indicating that there is no distortion in $\widehat{S}$. We note that (i) the sampling distribution of $\widehat{C}_R$ seems more tighter around zero than $\widehat{S}$, and (ii) the sampling distribution of $\widehat{S}$ is a little bit shifted to the right in the presence of coincidental outliers. This slight distortion, however, quickly disappears as the sample size increases up to $T=300,500$.

In addition to the graphical information provided by all the box-plots, we have also computed the simulated bias and mean-squared error (MSE) of the three correlation estimators ($\widehat{C},\widehat{C}_R,\widehat{S}$). To save space, we focus only on the case where two outliers occur at the same time (i.e., corresponding to Figures 4, 5 and 6) and the T-5 distribution case. The results are reported in Table 1. As shown in the table, both

[5] Some selected papers include Boyer et al. (1999), Longin and Solnik (2001), Ang and Chen (2002), and Patton (2006).

[6] We acknowledge that this link to the asymmetric correlation literature has been suggested by a referee.

[7] Suppose that, more realistically, the coincidental outliers are stochastic and occur due to strong tail dependence. If so, the population correlation measure might not be zero, since there is dependence in the tail. Hence, the large sample correlation coefficients such as shown in Figure 4 might not be entirely spurious, and can partly reflect the influence of the tail dependence on the population coefficient. However, even in this case, the population correlation may be an unattractive measure since it is so strongly influenced by the tail dependence. This shows that the population version of the proposed median-based measure may be more attractive, because it is robust to heavy tail-dependence. We thank the referee for bringing this point to our attention.

[8] We have also produced the results for the other cases (no outlier, a single outlier in x_t only, and outliers in x_t and y_t at different times), but do not report them in the paper because the results are fairly similar to the case of coincidental outliers. All the figures for these cases can be downloaded from http://web.yonsei.ac.kr/thkim/downloadable.html. Whenever some results are not provided in the subsequent discussion, the complete set of all results are also available at the same website.

$\widehat{C}_R$ and $\widehat{S}$ display substantially smaller bias and MSE than the conventional measure $\widehat{C}$ over all sample sizes and all generating distributions. When the sample size reaches 300 or 500, both bias and MSE are almost negligible for both $\widehat{C}_R$ and $\widehat{S}$ whereas they are still large for $\widehat{C}$. When $\widehat{C}_R$ and $\widehat{S}$ are compared against each other, the median-based method $\widehat{C}_R$ is marginally better than $\widehat{S}$ in terms of both bias and MSE.

We now turn to the non-zero correlation case. We set the true Pearson correlation coefficient (C) to be 0.5 in all generating distributions. The results for the conventional measure $\widehat{C}$ are shown in Figures 7-10. Unlike the previous zero correlation case, the dampening effect is absent both when there is a single outlier in x_t (Figure 8) and when there are outliers in x_t and y_t at different times (Figure 9). Figure 10 shows that the effect of coincidental outliers is again spectacular. On the other hand, such a large distortion is not present in both $\widehat{C}_R$ and $\widehat{S}$ as shown in Figures 11 and 12 in the case of coincidental outliers. However, we find that the center of the sampling distribution of $\widehat{C}_R$ is a bit smaller than the true Pearson value 0.5 whereas the sampling distribution of $\widehat{S}$ is well centered at the true Pearson value, except the T-1 case.[9] Moreover, the sampling distribution of $\widehat{S}$ seems to be tighter around its center than that of $\widehat{C}_R$, which means that $\widehat{S}$ must be better than $\widehat{C}_R$ in terms MSE. We note that the discovered 'downward bias' of $\widehat{C}_R$ shown in Figure 11 is present only because we compute the bias with respect to the true Pearson value C. If the bias is computed as the difference between $\widehat{C}_R - C_R$, then it can be easily shown that $\widehat{C}_R$ exhibits no bias.

In summary, the simulation results indicate that (i) the conventional Pearson measure is heavily affected by outliers, especially coincidental outliers, (ii) the median-based measure is marginally better than the Spearman measure if the true Pearson correlation coefficient is zero or close to zero, but (iii) the Spearman measure is clearly better than the median-based measure in terms of MSE when the true correlation coefficient is away from zero.

5 An Empirical Example

In this section we provide a small empirical application where we estimate the conventional measure, the median-based measure, and the Spearman measure for a pair of stock returns. The stocks considered in our application are Key Corp (Key) and National Oilwell Varco, Inc (NOV). These two companies are among those who experienced sharp declines in October 2008 (the subprime mortgage crisis) so that there are coincidental outliers in their returns. In fact, the loss during the same week of October 6, 2008 was -45.87% and -43.21% for the two companies, respectively. Another reason for selecting these two is that these companies are operating in financial and manufacturing sectors, respectively, so that it is expected *a priori* that their return series may not be strongly correlated. We have downloaded weekly data from Yahoo Finance and calculated returns. The sample period is from the week of 1/3/2005 to the week of 8/18/2014, which leaves us 502 weekly returns. Table 2 shows the estimation results. The first row reports the estimated values for the three measures $(\widehat{C}, \widehat{C}_R, \widehat{S})$, which are 0.50, 0.24, 0.34, respectively. The value 0.5 for the Pearson measure is somewhat high, which is contrary to our expectation. However,

[9] To explain the bias of $\widehat{C}_R$ shown in Figure 11, let us come back to the formula of the median-based measure in (9). For illustration, we focus on the $N(0, 1)$ case in which κ_x and κ_y are not very different from μ_x and μ_y, respectively. Therefore, abstracting from the denominator in the formula in (9), the difference between the median-based measure and the Pearson measure depends on how much skewness the distribution of the product term $(x_t - \kappa_x)(y_t - \kappa_y)$ possesses because they corresponds to the median and mean of the product term, respectively. An extreme case is $C = 1$, which makes the distribution of $(x_t - \kappa_x)(y_t - \kappa_y)$ very close to the chi-squared distribution with 1 degree of freedom $\chi^2(1)$. With a moderate degree of correlation, the distribution of the product term is always skewed either to the left or to the right although its degree of skewness is smaller than $\chi^2(1)$. Hence, the median-based measure is always smaller than the Pearson measure in absolute value. The median-based measure can be regarded as a shrinkage version of the Pearson measure, shrunken towards zero, in which the intensity of shrinkage is automatically adjusted by the degree of asymmetry of the underlying distributions.

when the two robust measures are employed, the degree of correlation is decreased. Next, we have dropped the coincidental outliers during the week of October 6, 2008 and recalculated the three measures, which are 0.45, 0.24, 0.34 reported in the second row of Table 2. Only the Pearson measure is reduced. In fact, the two companies have the second largest coincidental outliers during the week of November 24, 2008; a coincidental surge in their returns of 50.61% and 31.44% for the two, respectively. Hence, we have dropped these outliers as well for recalculation. The re-estimated values are shown in the third row of Table 2; 0.41, 0.24, 0.33. Again, the Pearson measure is further decreased whereas the two robust measures are qualitatively unchanged. As expected from the simulation results, the median-based measure underestimates compared to the Spearman measure in all cases.

6 Conclusion

In this paper, we show analytically and by simulations that the conventional measure of correlation is heavily influenced by the presence of outliers. Simulation results indicate that the distortion caused by outliers can be fairly large, especially when outliers are present in both variables simultaneously. In such a case, the entire sampling distribution of the correlation coefficient is far outside the true value of zero even when the two variables under consideration are truly independent. Hence, it is highly possible that researchers can wrongly conclude 100% of the time that the two variables are linearly related. Documenting such a spurious correlation phenomenon, we consider some robust alternative measures and compare their performance in the presence of such coincidental outliers by Monte Carlo simulations. The simulation experiments show that robust measures are not influenced by coincidental outliers. Based on simulations and empirical evidence, we recommend that the Spearman measure should be used when coincidental outliers are suspected.

References

[1] Ang, A., Chen, J., 2002. Asymmetric correlations of equity portfolios, Journal of Financial Economics 63, 443-494.

[2] Bonato, M., 2011. Robust estimation of skewness and kurtosis in distributions with infinite higher moments. Finance Research Letters 8, 77-87.

[3] Boyer, B.H., Gibson, M.S., Loretan, M., 1999. Pitfalls in tests for changes in correlations, International Finance Discussion Paper 597, Board of Governors of the Federal Reserve System.

[4] Ergun, A.T., 2011. Skewness and kurtosis persistence: Conventional vs. robust measures, Discussion paper.

[5] Huo, L., Kim, T.H., Kim, Y., 2012. Robust estimation of covariance and its application to portfolio optimization, Finance Research Letters 9, 121-134.

[6] Kim, T.H., White, H., 2004. On more robust estimation of skewness and kurtosis. Finance Research Letters 1, 56-73.

[7] Longin, F., Solnik, B., 2001. Extreme correlation of international equity markets. Journal of Finance 56, 649-676.

[8] Patton, A.J., 2006. Modelling asymmetric exchange rate dependence. International Economic Review 47, 527-556.

[9] Schweizer, B., Wolff, E.F., 1981. On nonparametric measures of dependence for random variables. Annals of Statistics 9, 879-885.

[10] Stigler, S.M., 1989. Francis Galton's account of he invention of correlation. Statistical Science 4, 73-79.

[11] White, H., Kim, T.H., Manganelli, S., 2010. Modeling autoregressive conditional skewness and kurtosis with multi-quantile CAViaR. In 'Volatility and Time Series Econometrics: Essays in Honour of Robert F. Engle', Oxford University Press.

Table 1. Simulated bias and MSE of the three estimators ($\hat{C}, \hat{C}_R, \hat{S}$) with coincidental outliers (for the T-5 distribution only)

	Sample Size (T)	Bias	MSE
$\hat{C}$	50	0.92	0.85
	100	0.87	0.76
	300	0.71	0.51
	500	0.60	0.36
$\hat{C}_R$	50	0.02	0.02
	100	0.01	0.01
	300	0.00	0.00
	500	0.00	0.00
$\hat{S}$	50	0.06	0.02
	100	0.03	0.01
	300	0.01	0.00
	500	0.00	0.00

Note: The results for the other generating distributions are available from the following website:
http://web.yonsei.ac.kr/thkim/downloadable.html.

Table 2. Estimated values for the three estimators ($\hat{C}, \hat{C}_R, \hat{S}$) using the return data on KeyCorp and National Oilwell Varco, Inc.

	$\hat{C}$	$\hat{C}_R$	$\hat{S}$
All observations included	0.50	0.24	0.34
The largest coincidental outliers (2008/10/6) deleted	0.45	0.24	0.34
The 2nd largest coincidental outliers (2008/11/24) deleted as well	0.41	0.24	0.33

Figure 1. Sampling distributions of $\hat{C}$ (Box-plots): no outlier case with $C = 0$

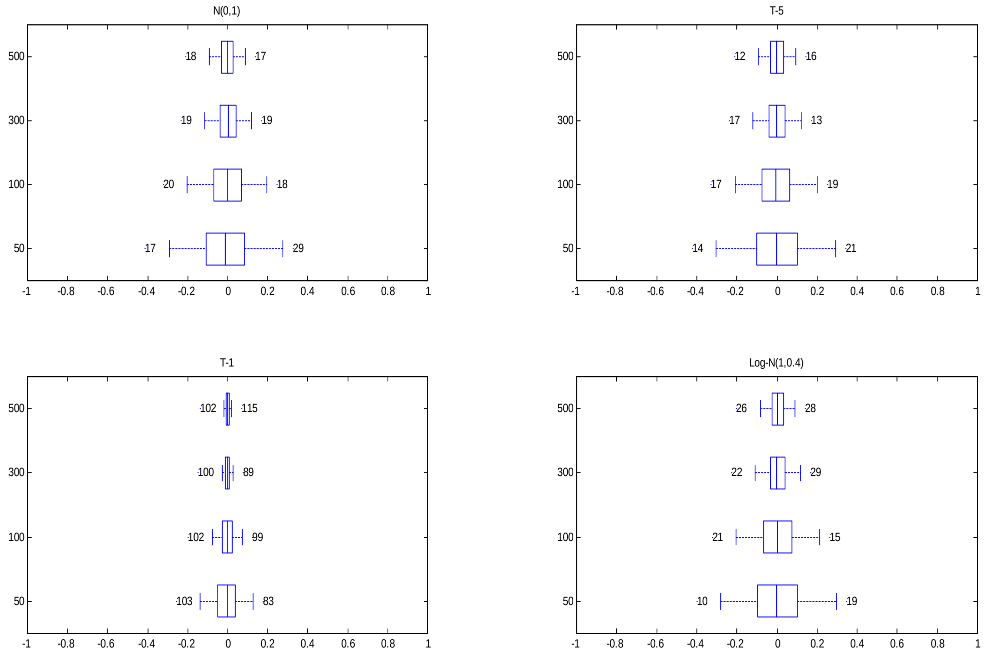


Figure 2. Sampling distributions of $\hat{C}$ (Box-plots): single outlier in x_t only with $C = 0$

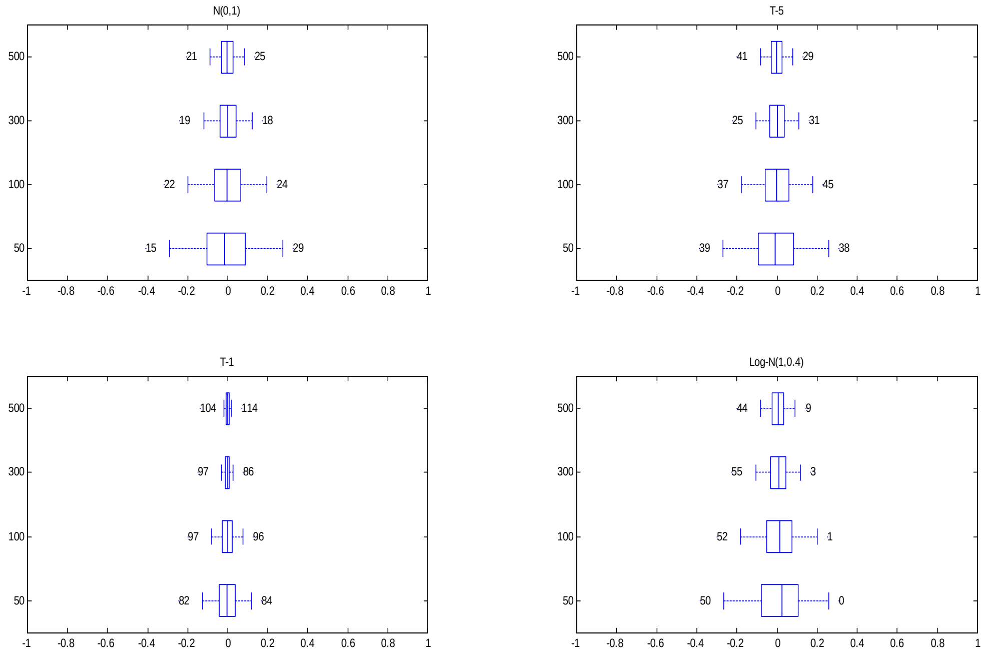


Figure 3. Sampling distributions of $\hat{C}$ (Box-plots): outliers in x_t and y_t at different times with $C = 0$

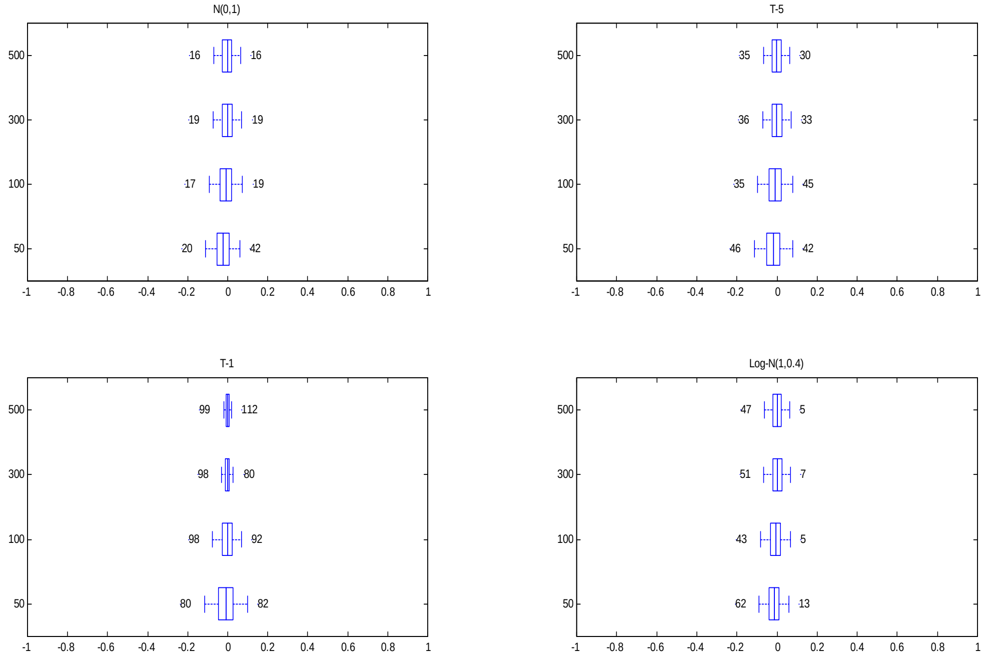


Figure 4. Sampling distributions of $\hat{C}$ (Box-plots): outliers in x_t and y_t at the same time with $C = 0$

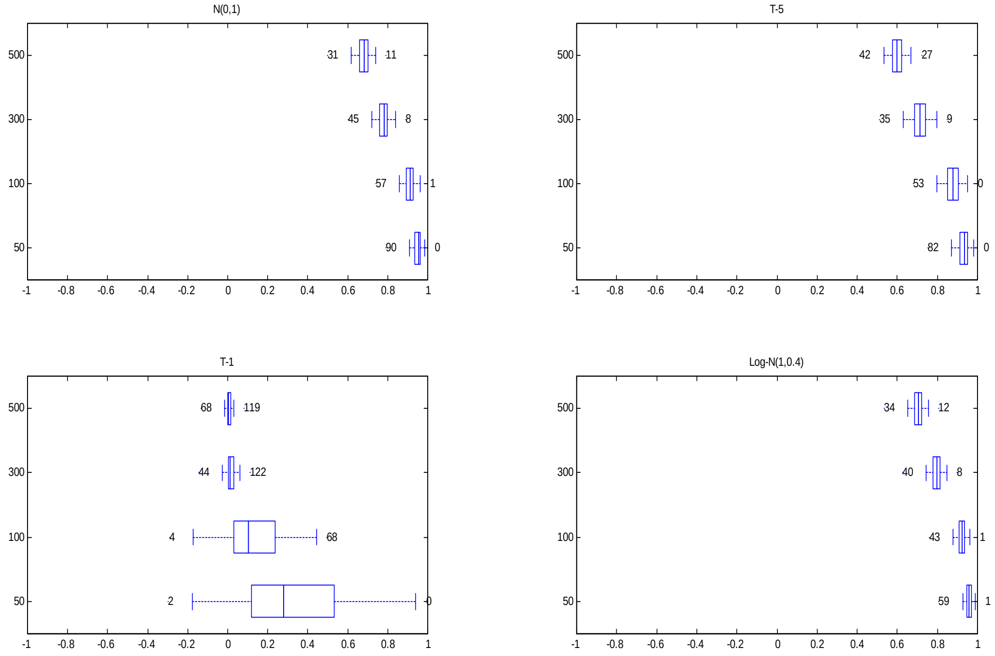


Figure 5. Sampling distributions of $\hat{C}_R$ (Box-plots): outliers in x_t and y_t at the same time with $C = 0$

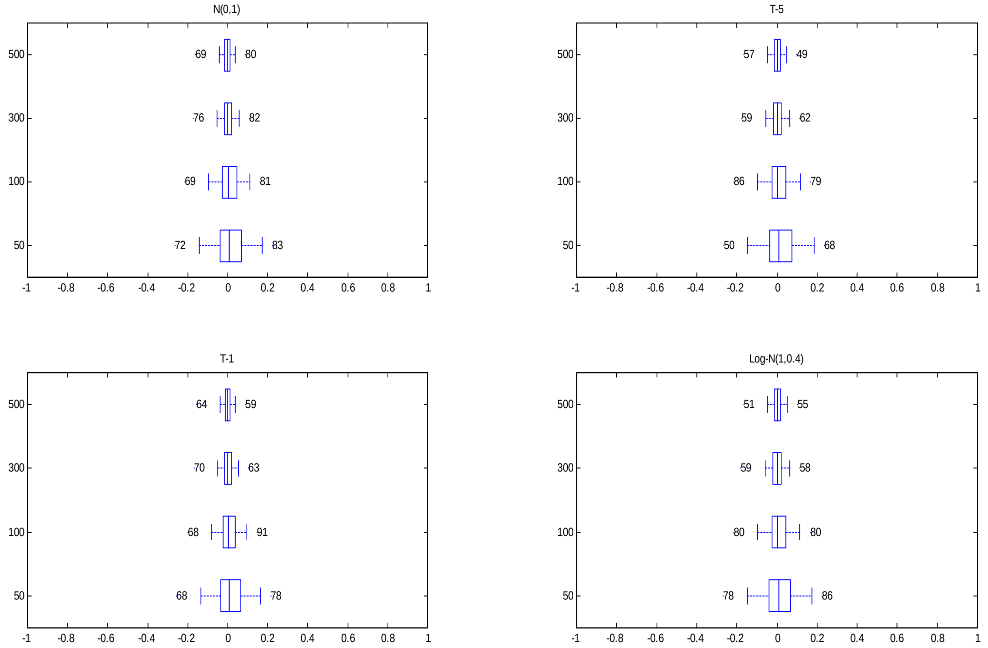


Figure 6. Sampling distributions of $\hat{S}$ (Box-plots): outliers in x_t and y_t at the same time with $C = 0$

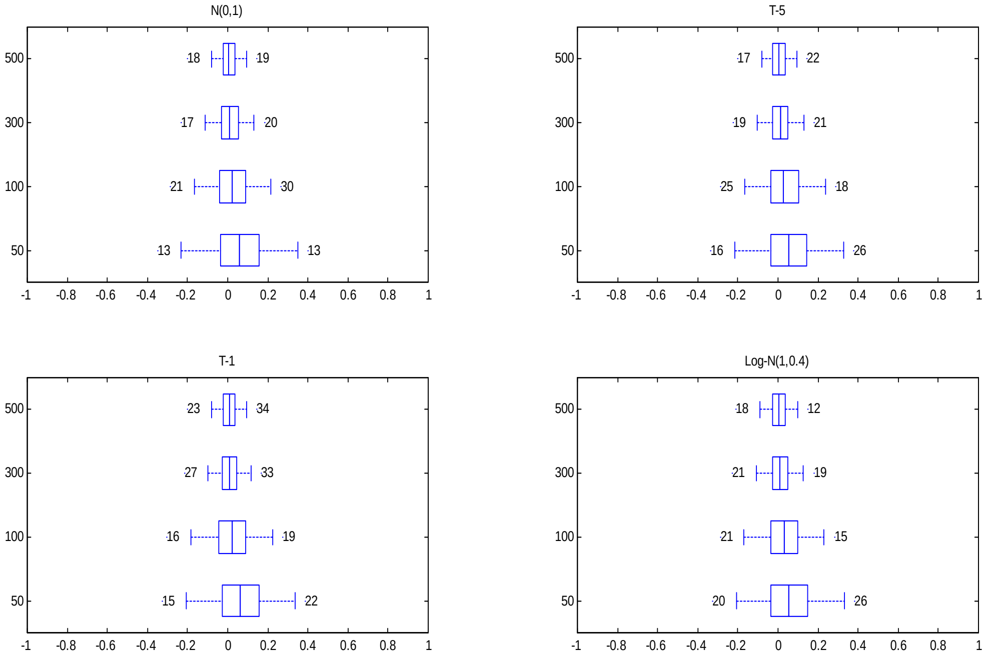


Figure 7. Sampling distributions of $\hat{C}$ (Box-plots): no outlier with $C = 0.5$

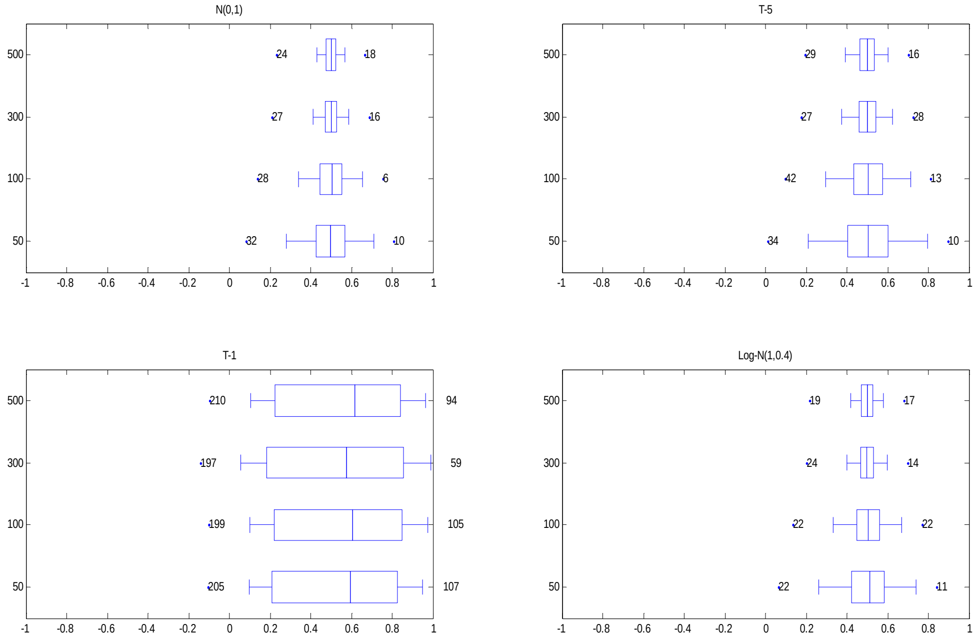


Figure 8. Sampling distributions of $\hat{C}$ (Box-plots): single outlier in x_t only with $C = 0.5$

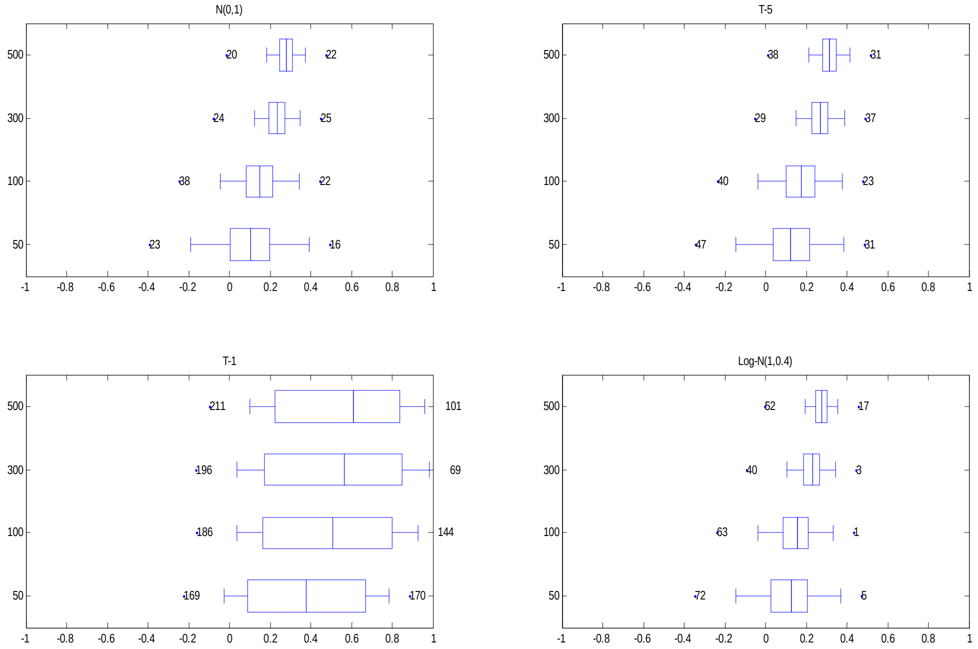


Figure 9. Sampling distributions of $\hat{C}$ (Box-plots): outliers at different times with $C = 0.5$

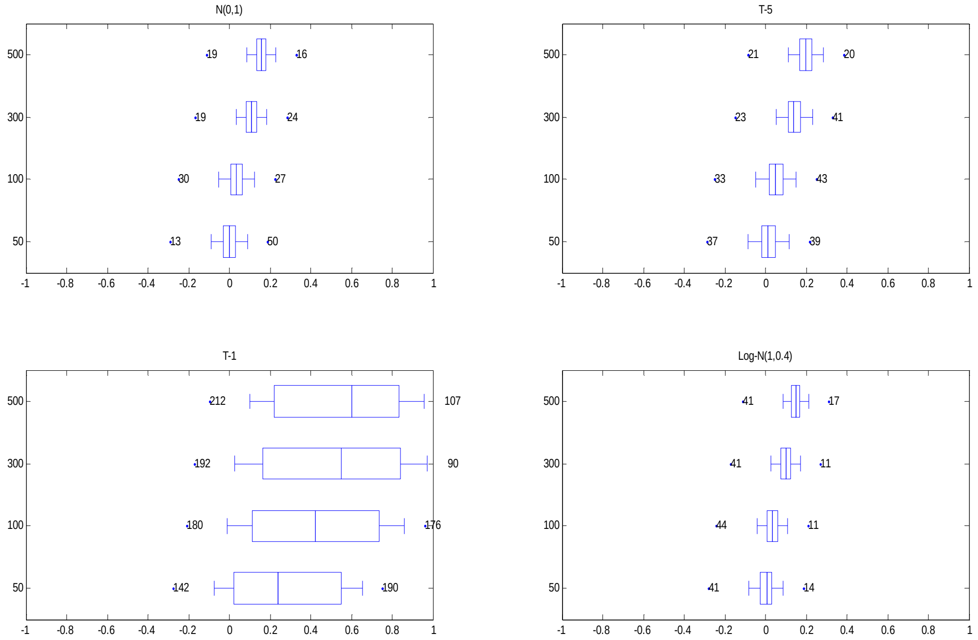


Figure 10. Sampling distributions of $\hat{C}$ (Box-plots): outliers at the same time with $C = 0.5$

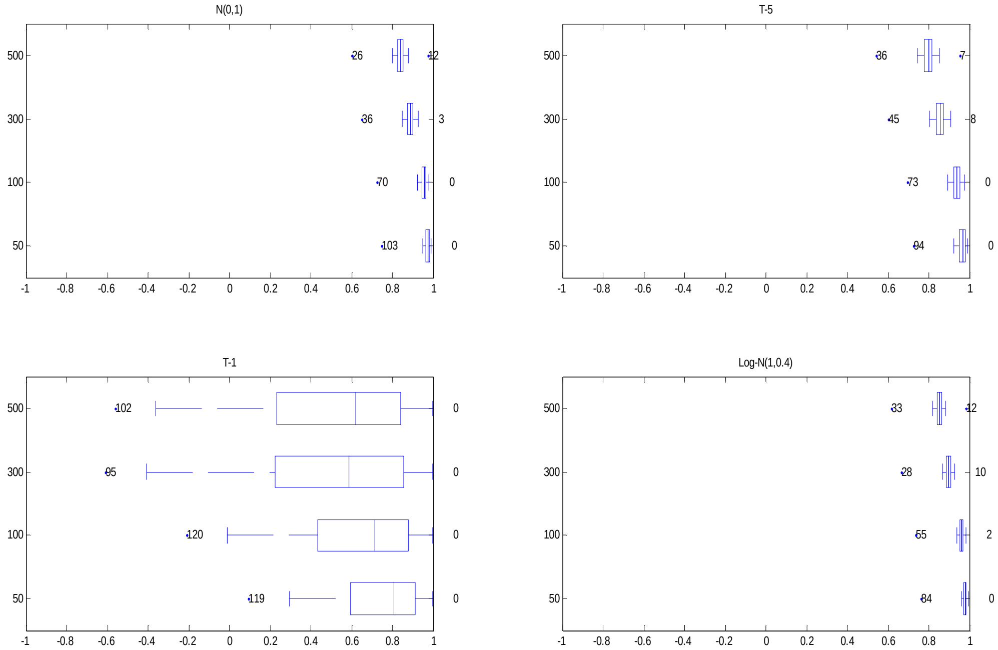


Figure 11. Sampling distributions of $\hat{C}_R$ (Box-plots): outliers at the same time with $C = 0.5$

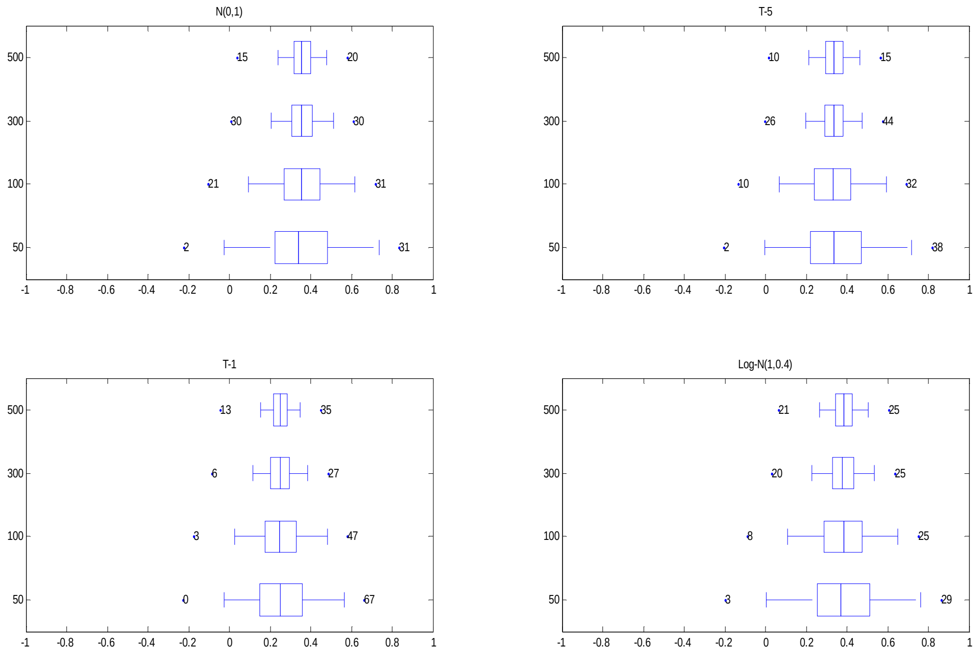
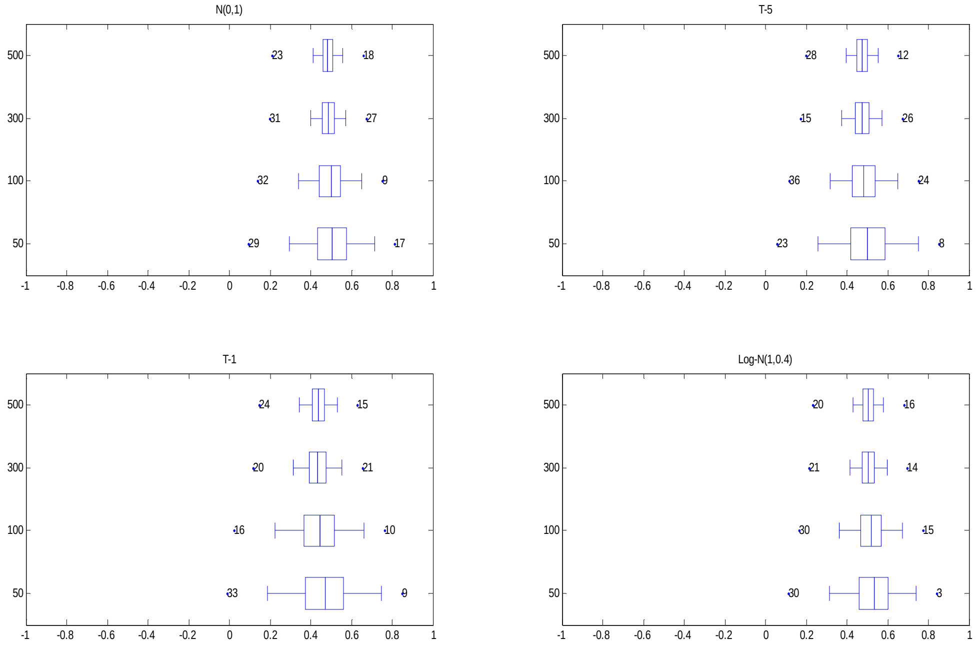


Figure 12. Sampling distributions of $\hat{S}$ (Box-plots): outliers at the same time with $C = 0.5$


N(0,1)
T-5
T-1
Log-N(1,0.4)