

Rock Creek Group Mutual Information Correlation

The Rock Creek Group
December 16, 2009

The Pearson correlation coefficient is a widely used measure of the correlation between two variables. However, it has its shortcomings, the most serious of which is that it assumes a linear relationship between the variables. In cases in which the variables have a non-linear relationship, e.g. the payoff of an option and the price of underlying asset, the correlation will be miscalculated. Moreover, even in cases in which the variables have a linear relationship, the correlation between the variables will likely be misstated in the presence of outliers. In this note, we propose another statistic, the RCG mutual information correlation coefficient, which is independent of the relationship between the variables. This measure is related to the entropy of the information.

I. Pearson Correlation

The linear correlation coefficient is calculated as

$$\rho = \frac{E((x - \bar{x})(y - \bar{y}))}{\sigma_x \sigma_y} \quad (1)$$

where x and y are two random variables, \bar{x} and \bar{y} are their mean values, and σ_x and σ_y are their standard deviation. This coefficient ρ is called Pearson correlation and captures the linear correlation. If the relation between x and y is non-linear, the Pearson correlation will misstate the relationship between x and y . The most common method for overcoming the non-linearity problem is using rank correlation coefficients. While rank correlation is a better measure for non-linear relationships, it is still an imprecise measure for capturing the relationship between x and y .

II. RCG Mutual Information Correlation

The RCG mutual information correlation coefficient is based on the concept of information entropy. Entropy is a measure of the uncertainty associated with a random variable. More specifically, it quantifies the missing information necessary to determine the value of a random variable. For example, if the underlying distribution of y is a Gaussian distribution, the missing information is clearly related to the standard deviation σ_y and the entropy in this case

$$H(y) \propto \ln \sigma_y \quad (2)$$

However, the entropy can be calculated for a wide range of distributions besides the Gaussian distribution. The entropy of a distribution with probability density $p(y)$ is defined as

$$H(y) \equiv - \int p(y) \ln p(y) dy \quad (3)$$

Given two random variables, we can calculate the conditional entropy

$$H(y|x) \equiv - \iint p_{x,y}(x,y) \ln \frac{p_{x,y}(x,y)}{p_x(x)} dx dy \quad (4)$$

If random variable x contains no information about random variable y , then the conditional entropy of $H(y|x)$ is identical to the entropy of the variable y .

$$H(y|x) = H(y) \quad (5)$$

This means knowing x will not reduce the amount of missing information on y . We define the mutual information of x and y to be

$$I(y,x) \equiv H(y) - H(y|x) = H(x) - H(x|y) \quad (6)$$

This measure quantifies the reduction in entropy (or missing information) of variable y because of variable x . In other words, the measure is the information common to the random variables x and y and is therefore related to the correlation between x and y . The mutual information is independent of the relationship between x and y , *i.e.* whether it is linear or nonlinear. In order to compare mutual information variable to the Pearson correlation coefficient, we define a measure called RCG mutual information correlation which is transformation of the mutual information variable $I(y,x)$. The RCG correlation coefficient will range between 0 and 1 and may be directly compared to the absolute value of the Pearson correlation coefficient.

Fig. 1 shows a simple Monte Carlo simulation test of the Pearson correlation, Spearman rank correlation and RCG mutual information correlation coefficients. It is clear that RCG mutual information correlation captures the non-linear relationship between an option payoff and the price of underlying asset, which the other two correlation measures do not.

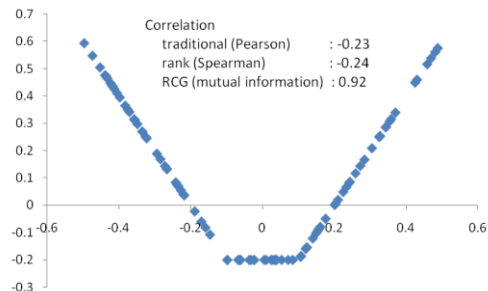


Figure 1: The correlation between the payoff of a strange option and the price of underlying asset. Since the payoff is determined by the price, the correlation should be one. However, due to the non-linearity, the traditional Pearson and Spearman rank correlations are low. The RCG mutual information correlation is very high. It is not exact one due to the limited number of data points.

III. Case Studies

We selected 825 funds from our database with more than 36 months of data over the past 5 years. The absolute values of the Pearson correlation coefficient and the RCG mutual information correlation were calculated for each of the funds against the S&P500 total return index. Fig.2 plots these two measures for all 825 funds. The difference between these two measures ranges from 0.47 to negative 0.40. Let us study these three extreme cases.

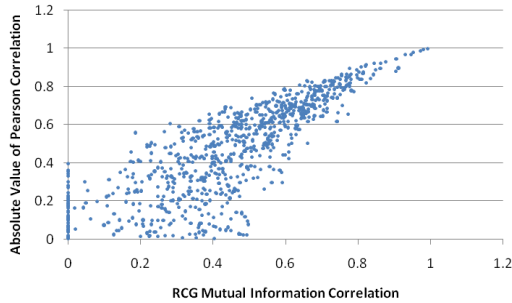


Figure 2: The absolute value of Pearson correlation and the RCG mutual information correlation.

In the first, (call it Fund 1) the Pearson correlation is 0.03 and the RCG mutual information correlation is 0.47. Fig. 3 plots the returns of Fund 1 and of the S&P500 total return index. A visual inspection suggests a clear relation between two return series. However, even though we use 5 years of data, a few outliers dominate and bias the Pearson correlation; however the RCG mutual information correlation is not sensitive to these outliers. When we use about 20 years of data for Fund 1 we observe that both correlation measures converge to 0.47.

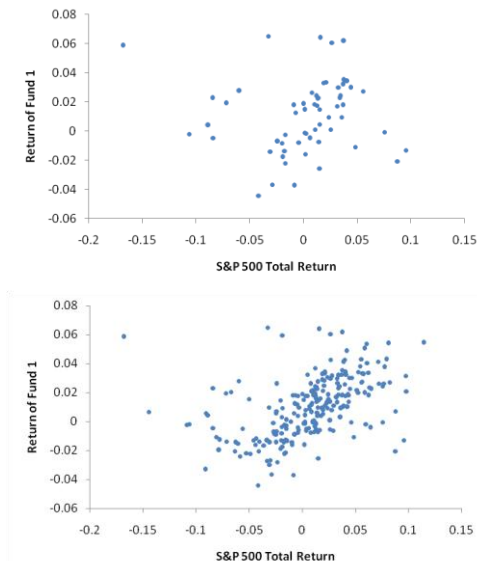


Figure 3: The return of Fund 1 (Y axis) with respect to the S&P 500 total return (X axis). The top plot contains the data from past 60 months. The Pearson correlation is 0.03 while the RCG mutual information ratio correlation is 0.47. The bottom plot contains the data from the past

277 months. The Pearson correlation is 0.47 and the RCG mutual information correlation is 0.68.

In the 2nd case (call it Fund 2) the data over the past 5 years give a Pearson correlation of 0.40 while the RCG mutual information correlation is 0. Fig. 4 plots the returns of Fund 2 and of the S&P 500 total return index. A non-zero Pearson “correlation” over 5 years is observed because of outliers. As before, the use of additional data mitigates the effect of outliers and both measures are consistent with no correlation.

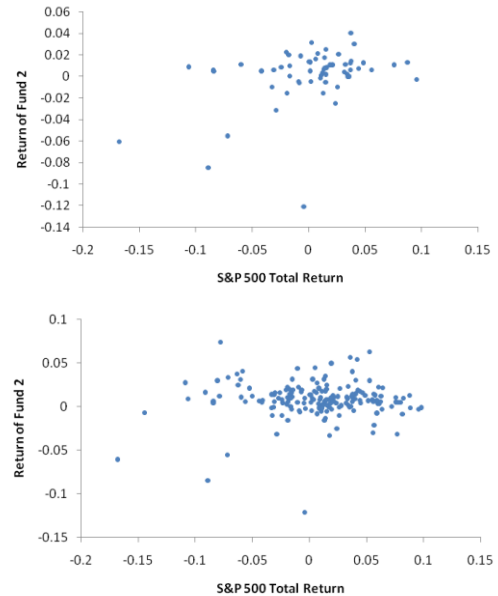


Figure 4: The return of Fund 2 (Y axis) with respect to the S&P 500 total return (X axis). The top plot contains the data from past 60 months. The Pearson correlation is 0.40 while the RCG mutual information ratio correlation is 0. The bottom plot contains the data from the past 172 months. The Pearson correlation is 0.05 and the RCG mutual information correlation is 0.09.

As is evident from the above two cases, the traditional Pearson correlation may misstate the ‘true’ correlation between two random variables in the presence of outliers and limited data sets. On the other side, the RCG mutual information correlation captures the correct relation between two series.

In the third case, (call it Fund 3) the Pearson correlation is -0.16 and the RCG mutual information correlation is 0.50. Fig. 5 plots the returns of Fund 3 and of the S&P 500 total return index. There are no obvious outliers in this case, and as such the difference in the correlation coefficients must be because of the non-linear nature of the relationship between the variables. To test this, we fit linear regressions between the Fund 3 returns and the S&P500 total returns, one for up markets and the other for down markets. Figure 5 shows a clear V shape relationship between the two variables. In other words, the relationship between the variables is non-linear and as such the Pearson correlation misstates the

correlation even as the RCG mutual information correlation will correctly estimate this non-linear relation.

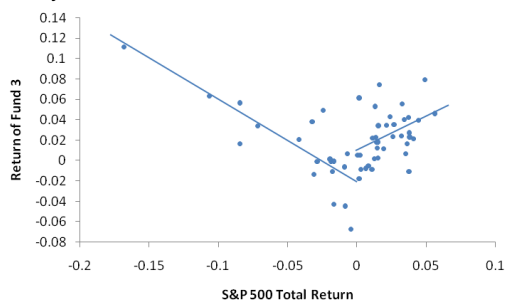


Figure 5: The return of Fund 3 (Y axis) with respect to the S&P 500 total return (X axis). Two solid lines are the linear regressions for up and down market respectively. The fund shows a V shape dependence on S&P 500 total return.

IV. Conclusion

The widely used Pearson correlation has significant shortcoming as it is not designed to capture non-linear correlations, and is extremely sensitive to outlier in limited data sets. Due to these shortcomings this measure can be misleading and result in drawing incorrect conclusions.

The RCG mutual information correlation is based on the entropy measurements. It can capture not only linear but also non-linear correlations. Moreover, it is insensitive to the existence of outliers and can correctly estimate the correlation with limited amount of data.