INSIGHT.DATA.CLARITY.

August 2015

THE MISLEADING NATURE OF CORRELATIONS

Executive summary

In this note we explain certain subtle features of calculating correlations between time-series. Correlation is a measure of linear co-movement, to be contrasted with the quadratic nature of risk. This can lead to misleading impressions arising from correlating two time-series. We show that the correlation of a manager with a benchmark leads to an estimate of the square root of how much exposure the manager has to the benchmark. We also show that an estimate of correlation with monthly data over five years has an associated error of 0.13, and therefore only a correlation of greater than 0.26 should be considered significantly greater than zero.

Contact details

Call us +33 1 49 49 59 49

Email us cfm@cfm.fr

Introduction

When comparing two return streams, investors generally calculate correlation coefficients to identify decorrelating and diversifying investments. Correlation calculations are ubiquitous enough to be included in any reasonable time-series analysis software package and are therefore often used blindly.

In discussing correlations we will also introduce the notion of exposure. For example if we combine two independent strategies, x and y, to give a combination $sum = x + y$, then the proportion of risk taken by x is represented by its β with the total[1] and that of y is similarly represented by its β with the total. The details of why exposure is defined in this way are described in the appendix.

In this note we will model real world return streams through the use of simple 'random walks' to illustrate a few counterintuitive results. A further appendix with a comprehensive derivation of the results is available upon request for the more mathematically inclined reader.

Numerical simulations – a pragmatic approach

In this section we will illustrate the power of using numerical methods to answer questions concerning the correlations between time-series. The following may be considered technical by some readers; it may be safely skipped in order to get to the key results. We first begin by introducing the basic tool of these simulations - the random walk. In order to keep things as simple as possible we will only study time-series with constant levels of risk and Sharpe ratio.

With this in mind, the simplest random walk for a price p can be written as follows:

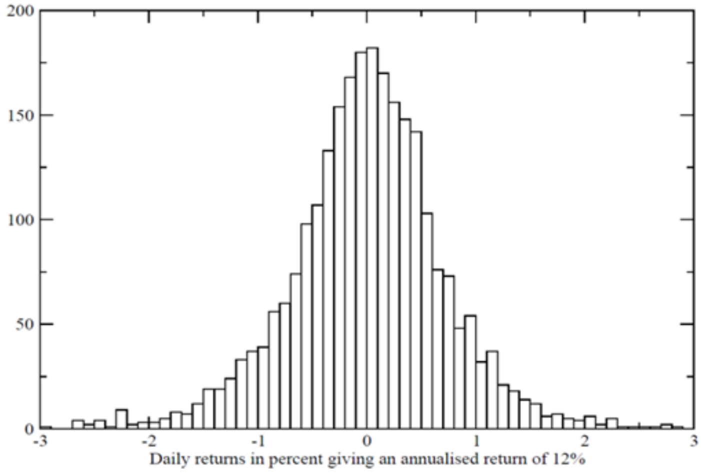

Figure 1: A histogram illustrating the bell shaped distribution of the random numbers used in the random walks. The random numbers are centred on zero and have tails that fit financial time-series well.

$$p_n = \sum_{n=0}^{N} (d + \eta_n)$$

Where n is the counter, say the days for a daily return and N is the total number of days in the time-series of returns. The η term is simply a zero mean noise term or random number generator with a bell shaped distribution that best models the returns of the investment strategy. A histogram of these random numbers can be seen in Figure 1 showing a distribution centred on zero with tails representative of financial returns[2]. The d term is a constant added to the unpredictable 'noise' η_n at every time step to generate a random walk with a 'drift,' or positive return. Figure 2 shows the results of generating random walks with Sharpe ratios of 0, 0.5 and 1 by varying the drift term to achieve the Sharpe ratio we require. Obviously, a Sharpe ratio of zero is generated by applying no drift term at all i.e. setting d to zero and allowing the zero mean of the η_n random numbers to generate a flat (on average) zero Sharpe random walk with a Sharpe ratio of zero.

We now have a framework within which to simulate many random walks with any particularly desired Sharpe ratio, each realisation being different due to the existence of the η_n term. The time-series in Figure 2 shows how these random walks resemble different return streams, such as investment indices or individual funds.

[1] The β of variables a with respect to b is defined as Covar(a,b)/Var(b) where Covar is the covariance between two variables Covar(a,b) $= \frac{1}{N}\sum_{n=0}^{N} a_n b_n$ while Var is simply the variance of a variable, more commonly known as the square of the standard deviation.

[2] The choice of the distribution of returns can change the results of the study. Here we use a Student's distribution with 4 degrees of freedom, a distribution which is naturally 'fat tailed' and fits financial time-series well. For the purpose of this short note, however, we will neglect the effects of these fat tails on the calculation of correlations. One could use the commonly known Gaussian distribution to achieve very similar results.

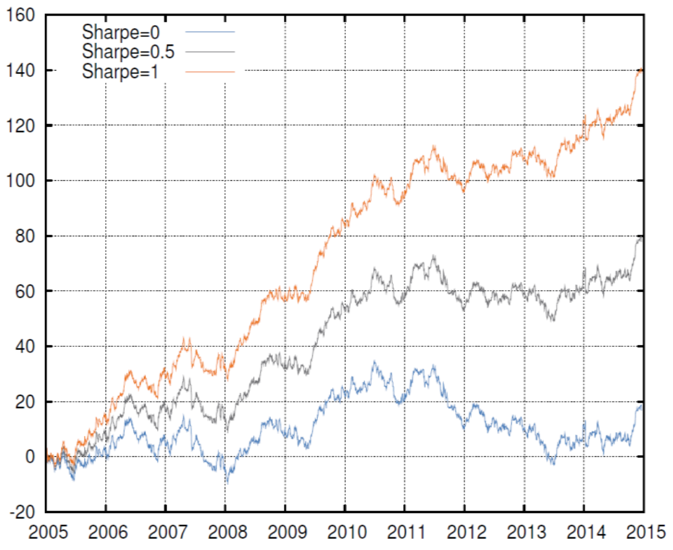

Figure 2: Random walks generated with three Sharpe ratios, illustrating how varying the d parameter allows us to easily change the drift and hence the Sharpe ratio.

Correlating two uncorrelated random walks with their sum

Let us imagine we have two time-series which are zero correlated, representing two different funds. These two time-series are shown in **Figure 3**. We have added a drift to get Sharpe ratios of one for each, and can now sum the two together. There is perhaps no surprise that the Sharpe ratio increases, showing the benefit of diversification, but let's now try to calculate the correlation of one of the strategies with the total. Intuitively one might expect that the correlation would be 50% due to the fact that we have 50% of each strategy in the time-series. In fact, the correlation turns out to be 71%! Correlating the sum of two time-series with either of the two strategies used in the sum gives us a higher correlation than the weight of the strategy within the mix. This could be considered a counterintuitive result. We will now show that correlation is always higher than exposure.

Correlation to evaluate a manager's exposure to a benchmark

We now turn our attention to another example. This time we have a manager with a small exposure to a well-known benchmark strategy, such as trend following, equity momentum, carry, value etc., but claiming he has decorrelated strategies running in parallel that make up the bulk of the risk of his returns. In order to estimate the contribution of a manager's return arising from a standard factor, an analyst may choose to correlate the benchmark or factor with the manager's returns. We can now use the example of the previous section (correlating the sum of two random walks with one of the two components) to illustrate how this can yield misleading results. We now allocate a proportion f of the benchmark strategy to the manager and combine it with $(1-f)$ of the uncorrelated non-benchmark strategy that the manager claims to be employing. Here we have a potential source of confusion as f does not reflect exposure, but it is instead the β of the strategy with respect to the total that is a true indicator of the risk taken by the strategy in the combination (please refer to the appendix for more detail on this point). We now have two time-series to correlate: the manager's returns $r_{man} = f r_{BM} + (1-f) r_{NBM}$ and the benchmark strategy r_{BM}, where r_{BM}, and r_{NBM} represent the returns for the benchmark and for the manager's decorrelated non-benchmark return streams respectively.

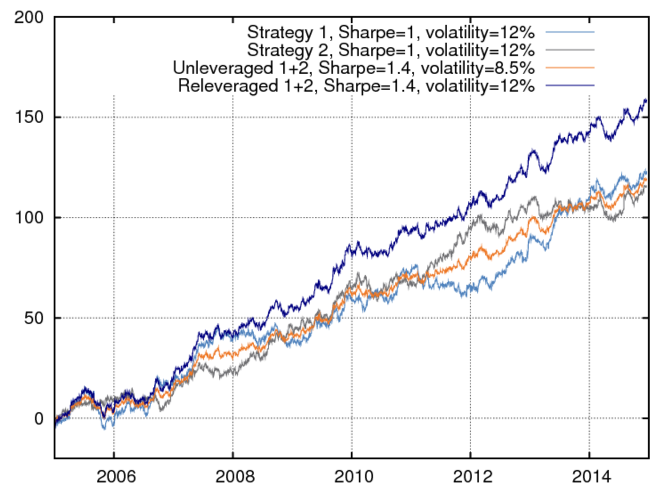

Figure 3: Two strategies, each with a Sharpe of 1, added together to illustrate the power of diversification. We first add each strategy with a weight of 1 half, thus obtaining the same level of drift but a lower volatility. We then leverage the volatility to be the same level as the two inputs, thus demonstrating that we reach a higher overall gain over the period. Correlating either of the two initial strategies with the sum gives a correlation of 71% rather than 50%, as naively expected.

Let's begin with the case of $f = 0.5$, which reproduces the result of the previous section, meaning a manager who has 50% of his risk allocated to a benchmark and 50% allocated to a non-benchmark strategy will correlate 71% with that benchmark strategy. Let's now try varying the weight f and observe how the correlation varies and, more interestingly, how the risk exposure to the benchmark varies. Because of the fact that risk sums quadratically, exposure to the benchmark strategy does not scale linearly with f (please see appendix for details). In **Figure 4** we plot the variation of the correlation and exposure as a function of f. One can see that the correlation does not follow the exposure, as stated, but is consistently above it. Correlation is, in fact, the square root of exposure. If we come back to the example of a 50/50 split between strategies giving a 71% correlation with the total, one can now observe that in fact the exposure of r_{man} to r_{BM} is $0.71^2 = 0.5$ which seems indeed logical. It suffices, therefore, in such situations to consider the square of the correlation as the best estimate of exposure to a particular strategy within a combination rather than just the correlation itself. We have shown this result empirically here but it can also be derived mathematically. Interested readers are invited to contact us for further details of the derivation.

The uncertainty on the measurement of correlation

Let us now turn our attention to the problem of the significance of a measurement. For correlations close to zero, the error on the measurement goes as $\sim 1/\sqrt{N}$ where N is the number of points used in the estimate[3]. If we assume that we are correlating managers with benchmarks using ~5 years of monthly data, then the error on the estimate is accordingly $\sim 1/\sqrt{12 \times 5} = 1/\sqrt{60} \sim 0.13$. Using daily data gives a far more significant result due to the fact that ~20 times more data is used in the estimate (as is the case in the analysis above). One needs to be careful in estimating correlations with monthly data where for a sample size of ~5 years, a correlation of 0.26 cannot (and should not) be considered positive (or negative!) with an acceptable level of significance.

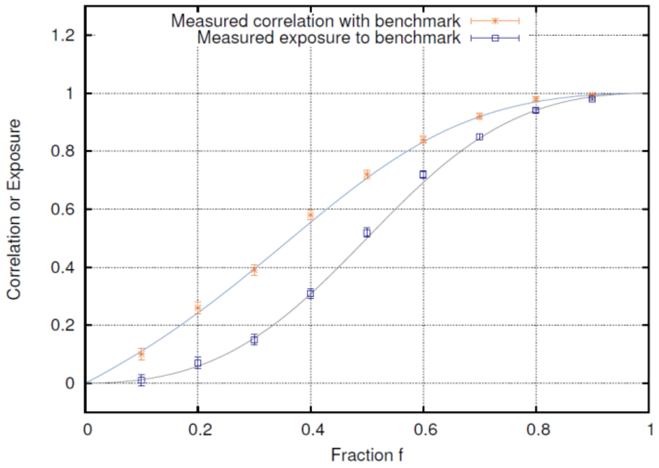

Figure 4: The plot shows the effect of varying the weight of the benchmark strategy that the manager is running (x-axis) against the corresponding correlation that the combination has with the benchmark and the exposure the combination has to the benchmark (y-axis). The parameter f is simply the weight allocated to the benchmark, not the proportion of risk in the combination. This 'exposure' is being encapsulated in the β (see text and appendix). Correlation is not the same as exposure, the two being related such that exposure is equal to the square of the correlation. The lines through the points are the result of an analytical solution to the problem, the details of which are available upon request.

Conclusions

When comparing a manager with a benchmark, correlation is not a good direct indicator of the exposure that the manager has to the benchmark. The square of the correlation is actually an estimate of the exposure the manager has to the benchmark, which can be very different to the correlation itself. One should also be aware of the fact that any correlation, especially using monthly data needs to be considered along with its statistical error. Using five years of monthly data means that one needs correlations of greater than 0.26 to be considered statistically significantly different to zero.

[3] The error is actually $\frac{1-\rho^2}{\sqrt{N}}$ for non-zero values of ρ

Important disclosures

ANY DESCRIPTION OR INFORMATION INVOLVING INVESTMENT PROCESS OR ALLOCATIONS IS PROVIDED FOR ILLUSTRATIONS PURPOSES ONLY.

ANY STATEMENTS REGARDING CORRELATIONS OR MODES OR OTHER SIMILAR STATEMENTS CONSTITUTE ONLY SUBJECTIVE VIEWS, ARE BASED UPON EXPECTATIONS OR BELIEFS, SHOULD NOT BE RELIED ON, ARE SUBJECT TO CHANGE DUE TO A VARIETY OF FACTORS, INCLUDING FLUCTUATING MARKET CONDITIONS, AND INVOLVE INHERENT RISKS AND UNCERTAINTIES, BOTH GENERAL AND SPECIFIC, MANY OF WHICH CANNOT BE PREDICTED OR QUANTIFIED AND ARE BEYOND CFM'S CONTROL. FUTURE EVIDENCE AND ACTUAL RESULTS COULD DIFFER MATERIALLY FROM THOSE SET FORTH, CONTEMPLATED BY OR UNDERLYING THESE STATEMENTS.

Appendix

The correlation ρ generally takes the form:

1. $\rho = \frac{\text{Covar}(r_x, r_y)}{\sigma_x \, \sigma_y}$

where x and y represent two investments, r is the return and σ is the standard deviation or volatility of the investment. The 'Covar' represents covariance which is an averaged quantity over the two variables:

2. $\text{Covar}(r_x, r_y) = 1/N \sum_{n=0}^{N} (r_x(n) - \overline{r_x})(r_y(n) - \overline{r_y})$

where n is the counter, say the days for a daily return and N is the total number of days in the time-series of returns. $\bar{r}_x$ and $\bar{r}_y$ are the means of the returns r_x and r_y respectively.

For the purpose of the note, we would also like to introduce a couple of further ideas concerning risk. Let us take the example of two uncorrelated strategies x and y that are combined together to give a strategy $sum = x + y$. Let us now employ a weighting scheme allocating say 0.2 to x and 0.8 to y and write the risk of the strategy sum as follows:

3. $\sigma_{sum}^2 = 0.2^2\sigma_x^2 + 0.8^2\sigma_y^2 + 2\rho 0.2\sigma_x 0.8\sigma_y$

Note that ρ is the correlation between the strategies that we have set at zero and therefore we can neglect the last term. What one can see immediately is that with a portfolio weight of 0.2 we do not have a contribution to the total risk of 20%. In fact the contribution, because it is squared is much smaller at $0.2^2 = 0.04$ compared to the weighting of the strategy y which takes a weight of $0.8^2 = 0.64$ in the calculation of the variance[4]. If we set the volatility of each to be $\sigma_x = \sigma_y$ in order to simplify further then on a stand-alone basis with such a weighting the risk of strategy x is indeed 0.2 and that of strategy y is 0.8. However, it is clear that in the combination the contribution to the risk of strategy x is much smaller than 20%. So, the question one should ask is how much of the risk is being explained by x and y in the total?

Let's now therefore try to find a definition for exposure. Consider instead a regression analysis similar to that used in the CAPM, regressing the returns r_x of the strategy x and the returns r_y of strategy y on the returns r_{sum} of the combined strategy. We can define a weight f as the allocation weight to strategy x (this corresponds to the 0.2 in the above example) and a weight $1 - f$ to the strategy y (equally this corresponds to the 0.8 in the above example), such that the sum of these portfolio weights is equal to 1. As above however, we note that the portfolio weights are not a measure of the contribution of risk of x and y in the sum. Now, analagously to the CAPM we regress the strategy returns on $r_{sum} = fr_x + (1-f)r_y$ to give:

4. $fr_x = \beta_x r_{sum} + \text{'unexplained'}$

5. $(1-f)r_y = \beta_y r_{sum} + \text{'unexplained'}$

Given that f is known and that sum is known to only be composed of r_x and r_y then the unexplained part of the regression collapses to zero and we can explain sum fully with x and y. The βs also conveniently add linearly to 1 and are proportional to the amount of risk carried by the corresponding strategy as a fraction of the total. In such a case the βs are calculated as follows:

6. $\beta_x = \frac{Covar(fr_x, r_{sum})}{Var(r_{sum})}$

7. $\beta_y = \frac{Covar((1-f)r_x, r_{sum})}{Var(r_{sum})}$

and assuming again that $\sigma_x = \sigma_y = 1$ the βs can be written as:

8. $\beta_x = \frac{f^2}{f^2 + (1-f)^2}$

9. $\beta_y = \frac{[1-f]^2}{f^2 + (1-f)^2}$

Interested readers are again asked to contact us for the derivation of the above analytical solutions which are used in Figure 4 in the paper. So, 'Covar' is covariance as previously defined and the unexplained noise terms in the regression are fully explained by the fit. Let us now once more draw on the analogy with the CAPM in terms of interpreting the βs, looking at equations (4) and (5) one sees clearly that on average $fr_x = \beta_x r_{sum}$ and (1-f)r_y = $\beta_y r_{sum}$ meaning that for a given return of the sum r_{sum}, the amount by which the strategies x and y move relative to sum is encapsulated by the βs. Since the two βs sum to 1, they represent how much of the risk (or strictly speaking the variance) is being explained by each strategy in the total. As mentioned previously, this relates back to the CAPM. When we regress stock returns r_i over the index returns r_I we obtain the following

10. $r_i = \beta_i r_I + \text{'unexplained'}$

and β is interpreted as the exposure one gets to the index by holding the stock. If the β is 2 and the index goes up by 1%, we would earn 2% on average from our investment. Here we are doing much the same thing except the index is now the sum of two investments rather than an arbitrarily weighted index and the returns we are regressing are those of the strategies we used to construct the sum.

[4] variance is just volatility squared

CFM has pioneered and applied an academic and scientific approach to financial markets, creating award winning strategies and a market leading investment management firm.

Contact us

Capital Fund Management S.A.
23, rue de l'Université,
75007 Paris, France
T +33 1 49 49 59 49
E cfm@cfm.fr

CFM International Inc.
The Chrysler Building, 405 Lexington Avenue - 55th Fl.,
New York, NY, 10174, U.S.A
T +1 646 957 8018
E cfm@cfm.fr

CFM Asia KK
9F Marunouchi Building, 2-4-1, Marunouchi, Chiyoda-Ku,
100-6309 Tokyo, Japan
T +81 3 5219 6180
E cfm@cfm.fr

Capital Fund Management LLP
64 St James's Street, London
SW1A 1NF, UK
T +44 207 659 9750
E cfm@cfm.fr