Correlation measures are often major drivers of value-at-risk. *Brett Humphreys* and *Eric Raleigh* review the assumptions associated with calculating correlation

Valid assumptions required: calculating correlation

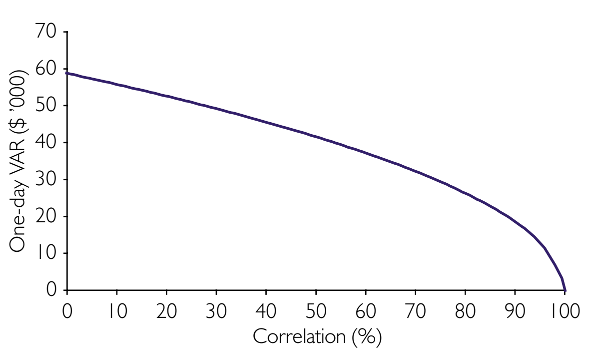
★ Few market participants have a simple long or short position. Instead, native long or short positions are hedged, sometimes with perfectly matching instruments but more frequently with related assets. The effectiveness of these hedges is driven by the strength of correlation between two assets. Because of this, value-at-risk and changes in VAR are fundamentally dependent upon correlation and changes in correlation.

The statistical definition of correlation is the covariance between two variables divided by the standard deviation of each variable:

$$\rho = \frac{\text{cov}(XY)}{\sigma_x \sigma_y}$$

By definition, correlation must be between minus one and one. Because of this standardisation, we tend to discuss the relationship between two series in terms of correlation rather than covariance. Of course, even this calculation requires many assumptions.

To see the impact of correlation on VAR consider the following portfolio – long $1 million of asset 1 and short $1 million of asset 2. Assume both assets have an annual volatility of 40%. Figure 1 shows how the VAR of this position changes as the correlation between asset 1 and asset 2 changes. The first thing we can see from this figure is that if the correlation is less than 50%, the risk of the portfolio is greater than the risk of one leg. In other words, we need at least a positive 50% correlation to use an asset as a hedge. If we want to cut the risk in half, we need at least an 85% correlation between the two assets.[1]



F1. Sensitivity of VAR to correlation

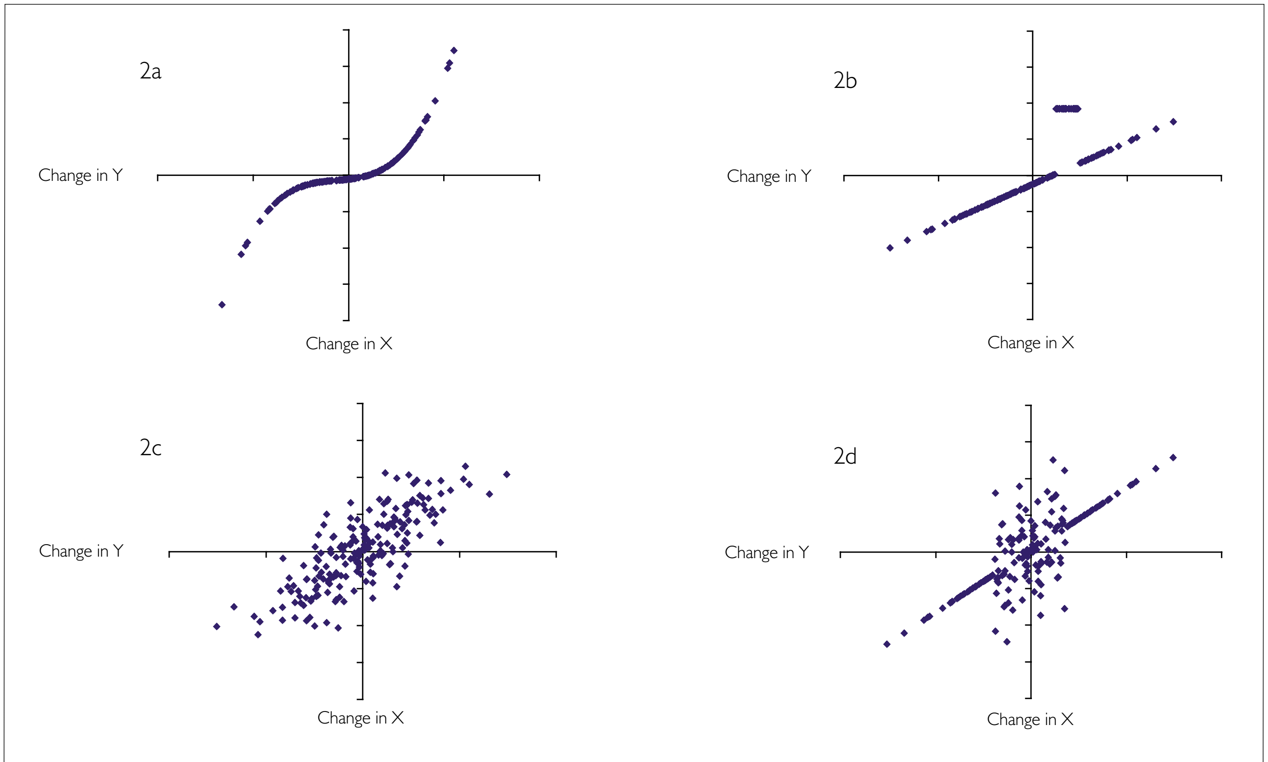
The figure also shows that the relationship between VAR and correlation is not linear. In fact, at extremely high correlations (greater than 95%) slight correlation changes can have significant effects on VAR. At first glance, we might say there is little difference between a correlation of 99% and 96%. However, the portfolio with a 96% correlation has a VAR twice as large as that based upon a 99% correlation. When we consider that there is uncertainty in our estimates of correlation, we can easily see how slight variations can have a significant impact on VAR.

One issue is that frequently we calculate a correlation measure and think we understand what it means. For example, assume we are told that two assets have a correlation of 80%. Mentally, we have an image of what we expect the joint distribution of these two assets to look like. Examine the four graphs in figure 2. Each of these data sets has a correlation of 80%. In addition, each series has the same mean and standard deviation. In reality, when we are told something has a correlation of 80% we assume a relationship like figure 2c – in other words, we assume underlying normal distributions of the data. However, that might not always be the case. We must remember that correlation is simply a statistical measure of joint dispersion between two distributions and can be calculated on any distribution.

The problem as usual comes down to assumptions. As correlation calculations are fundamentally related to volatility calculation, correlation embeds those assumptions:

- A functional form (geometric Brownian motion, mean reversion, etc) of the two underlying price processes.
- The historical data set is appropriate for predicting future correlations.

1. It is worth noting that hedge effectiveness testing for hedge accounting treatment occasionally refers to a 'correlation' between an asset and its hedge of 0.8 and 1.2. This is not a statistical measure of correlation. It is instead the predicted change in hedge value given a specific change in asset value. However, the true statistical correlation is extremely important to ensure that the hedge will remain effective through time. In a regression analysis, correlation is equivalent to regression $\sqrt{R^2}$. Without a high R^2, the hedge will not effectively mitigate risk and large portions of profit and loss will be deemed ineffective



F2. Series with 80% correlation
All data sets have the same mean and standard deviation

• The underlying distributions are normally distributed.

In addition, the standard correlation calculation includes two other important assumptions:

• The relationship between the two series is linear.
• The relationship between the two series is constant.

A linear relationship

It is important to realise that a correlation calculation fits a line to a data set. The closer the data is to that line, the higher the calculated correlation. The problem is that the relationship may be more complex than this assumption implies. In figure 2a, we see that the change in *Y* can be perfectly predicted given the change in *X*. However, the change is not a simple linear relationship. In this situation, it should be possible to perfectly hedge an exposure to *X* with *Y* even though the naive correlation calculation is 80%.

While figure 2a represents a hypothetical example, there are more realistic scenarios. Consider the relationship between changes in implied volatility and price. We would expect that both significant increases and decreases in price would lead to increases in implied volatility. In a previous analysis examining the correlation between changes in implied volatility and prices, we found that changes in implied volatilities had a +22% correlation with positive changes in price and a –28% correlation with negative changes in price.[2] In this case, ignoring the more complex dynamics between volatility and price would lead to a mis-estimation of the correlation and potentially the risk of a portfolio containing options.

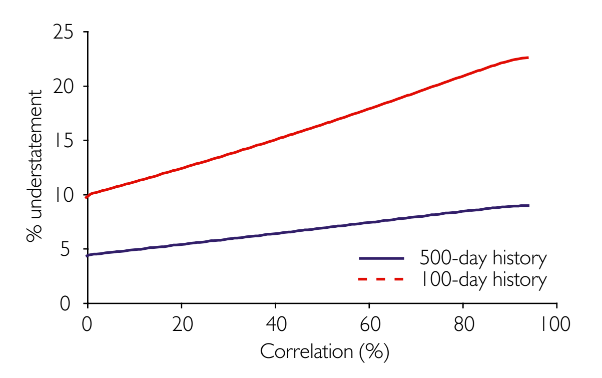
A constant relationship

Another significant assumption is that the relationship between the two series is constant. In reality, this may be violated in multiple ways, but two of the most common are that the relationship may be structural or that the relationship may differ based upon the size of the price changes (figure 2d is a theoretical example of this).

In the power and gas markets, a common structural impact is seasonal correlation. Seasonal correlations may arise between two connected regions, especially when the connection becomes constrained. In this situation, when the transportation is unconstrained, the price differential should be limited by the transportation costs and price changes in each region should be highly correlated. However, at times when transportation becomes constrained, the prices in each region can move independently and the correlation declines. When transportation capacity becomes systematically constrained at certain times of year then this will lead to seasonal correlations.

In the oil markets, we can see structural correlation patterns

2. Viva Lost Vegas, Energy Risk, August 2004



F3. Potential VAR understatement due to correlation uncertainty

based upon backwardation or contango. In a full contango market, there is arbitrage between contracts that are close to delivery and contracts that are further from delivery. Because of this cash and carry arbitrage, movements of all points of the forward curve will be highly correlated. In contrast, in a backwardated market (one in which the price of a contract close to delivery is higher than a contract for delivery further in the future) shortage of physical product constrains arbitrage. The result is that the front of the curve can move relatively independently of the back of the curve and correlation declines.

Correlation may also depend on the size of the price changes we evaluate. Consider two related markets with relatively high transportation costs. The price of each market may vary up to a certain point, but once it crosses that threshold arbitrage constrains further movement. The result is a fairly common phenomenon: a low correlation for small changes and a high correlation for large price changes. For example, we calculate a 92% correlation between WTI crude and Brent crude across all observations. But if we segregate the data by large returns (> 1.5 standard deviations), the correlation rises to 98%.

Accuracy of correlation measure

We must also always remember that correlation is only an estimate of the true population correlation. However, because correlation is bounded, the distribution around the estimate is not normal. We can, however, use Fisher's z to calculate the significance of correlation estimates.[3] For example, if we use 100 observations and calculate a correlation of 80%, the true correlation could be anywhere between 86% and 72%.

How does this uncertainty affect our VAR measure? Let us return to our initial example of long $1 million of asset 1 and short $1 million of asset 2. Figure 3 shows the potential VAR understatement due to correlation uncertainty based on a correlation derived from 100 days of history and a correlation derived from 500 days of history. If we use 100 days of history, we see that we may be understating VAR by as much as 20%. If we use 500 days of history (assuming of course that correlation has not changed over that period) we may still be understating VAR by 5%. Given the sensitivity to correlation, we must do everything we can to avoid introducing any problems in our calculation. Even a few bad or stale data points can quickly bias a correlation calculation (see figure 2b), especially if the correlation is close to one. Plotting data is an excellent way to identify outliers and stale prices.

3. A Pearson's correlation coefficient is non-normally distributed. To calculate confidence intervals we calculate Fisher's z. $z = 0.5\ln((1 + r)/(1 - r))$. z is distributed normally with variance $1/(N - 3)$. r is the correlation estimate and N is the number of observations used to calculate correlation

Helpful hints

- Correlation should only be calculated on price returns.
- Correlation calculation is very sensitive to bad data points. The underlying data should always be carefully checked for outliers.
- A first approximation if correlation is significantly different from zero is to compare it with $2/\sqrt{n}$. For example, a correlation calculated from 100 observations would have to be greater than 20% for us to say it was statistically different from zero.
- The potential VAR understatement on a spread position is roughly $2/\sqrt{n}$, where n is the number of observations used to calculate the correlation.

Conclusions

Risk managers need to know their portfolio's exposure to correlation. With a diversified portfolio, we need to worry about strengthening of correlation, as all asset prices will have a greater tendency to move together, reducing the diversification benefit. With a hedged portfolio we need to be more concerned about a weakening of correlation, where our hedges will become less effective.

While we like to believe we understand correlation, we frequently ignore the assumptions we impose on a simple correlation measure. However, given the importance of correlation to portfolio risk measures, it is crucial that we revisit and question these assumptions. This process will increase our own understanding and, eventually, lead to better metrics. ER

Assumptions	Implications
The relationship between the two series is linear, that is, $\Delta Y = a + \beta\Delta X$	If the relationship is more complicated then a simple line, correlation will not pick that up
Estimate correlation from historical data set	Historical observations are all drawn from the same distribution and are indicative of the future distribution. There exists no structural difference between time periods (seasonal correlation) or by size of the return.
Note: we have focused on assumptions specific to correlation calculations. Remember that all assumptions related to volatility calculation also apply.	