portfolio construction forum

Blinded by optimism

Mark Roulston & David Hand | Winton Capital Management | 22 August 2016

Overfitting is a well-known problem and one would expect clever statisticians and scientists not to succumb to its temptations. But what amounts to overfitting can occur more subtly through the collective behaviour of many individuals within an organization or across many organisations. This paper describes how such "meta-overfitting" may be endemic in finance as well as other fields of research.

1. INTRODUCTION

Mark Twain is often quoted as saying, “It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so."[1] A perennial problem in both science and finance is knowing what is signal – systematic and predictable effects that are likely to persist – and what is noise – random fluctuations that one should not expect to persist. A common pitfall is to misidentify noise as signal and bet that it will persist. When this is done, incorrect conclusions can lead to disastrous consequences – such as the effectiveness of a new drug, or the profitability of an investment.

There have been articles in *The Economist*, *Nature*, and elsewhere, which suggest that this problem is more common than has been generally appreciated in the scientific literature[2] [3] [4] [5]. The issue addressed is that of publication bias – a tendency for papers with a positive detection of some effect to get published, and papers reporting a null result not to be published. When this bias is compounded by the confusion of noise with signal, so that a false positive occurs, the inevitable result is that many published conclusions are actually wrong.

This paper highlights the fact that these issues are also relevant to the world of finance, with equally serious consequences. Whether it is investors having to wade through advertising literature from funds and investment vehicles, or "quants" inside banks having to pick and weight strategies, there is the potential for noise to get confused with signal. As with publication bias, when one selectively picks apparently well performing investments based on noisy estimates of their true performance then they may be overly optimistic about what to expect in the future, and therefore make poor investment decisions.

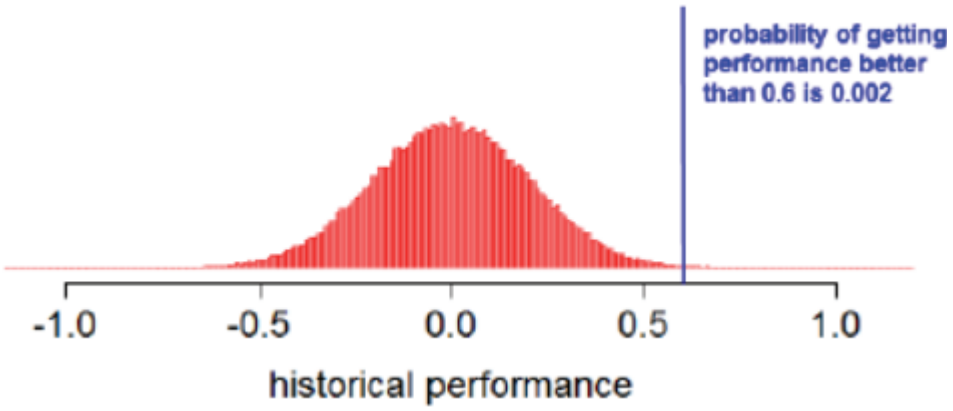
Noise can be confused with signal in many ways. Some of which are well known and relatively easy to avoid while others are more insidious and present a threat to the reliability of whole areas of scientific and financial research. Section 2 introduces some

basic concepts and terms. Section 3 discusses ways in which one can be misled, before going through possible solutions in Section 4 and a concluding discussion in Section 5.

2. HYPOTHESIS TESTING & "p-VALUES"

Let's begin with two workhorses of statistics: significance testing and p-values. In many areas of science, "null hypothesis significance testing" has become the standard framework for testing hypotheses. This framework starts by specifying a null hypothesis. You can think of this as effectively what we are going to believe unless there is strong evidence against it. Data is then condensed into a statistic, and its p-value is determined. This is the probability of getting a value of the statistic as extreme or more extreme than that observed if the null hypothesis is true. A small p-value means that either a very unlikely event has occurred, or the null hypothesis is false. Small enough, and we are inclined to reject the null hypothesis.

Figure 1: The performance distribution, as measured by the Sharpe ratio, of individual skill-free trading strategies, over 23 years. In this example, the probability of obtaining a Sharpe ratio better than 0.6 is only 0.002. This is the p-value.



The null hypothesis is analogous to the presumption of innocence in a criminal court. In finance, an appropriate null hypothesis is often the Efficient Market Hypothesis, the primary consequence of which is that price changes are not predictable. The statistic often considered is the historical performance of a trading strategy, which is expected to be zero in the case of efficient markets. If the p-value – the probability of obtaining the observed historical trading performance in efficient markets due to chance alone – is sufficiently small, we can feel justified in rejecting the null hypothesis, and infer that markets are probably not efficient and adopt an alternative explanation instead.

As an example, suppose a researcher presents a trading strategy that has a historic Sharpe ratio[6] of 0.6. The distribution of Sharpe ratios from skill-free strategies that buy and sell entirely at random over the same time period is calculated. This is the red distribution in Figure 1. Only 0.2% of these random strategies have Sharpe ratios better than 0.6. That means there's a probability (a p-value) of only 0.002 of getting such a high Sharpe ratio if the strategy has no skill. It is possible to conclude with some confidence that the new strategy "works".

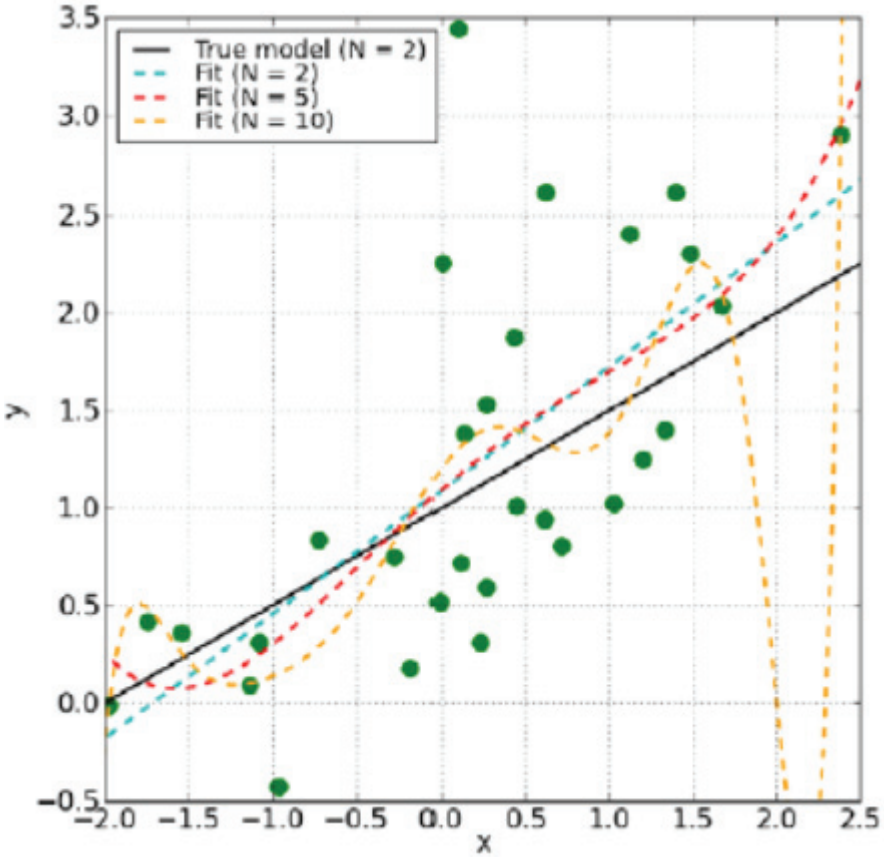
3. HOW TO CONFUSE SIGNAL AND NOISE

Various research practices can lead to unreliable results, and overconfidence in rejecting a null hypothesis. The root cause is often similar but the label given to the bias depends on the context.

3.1 Overfitting

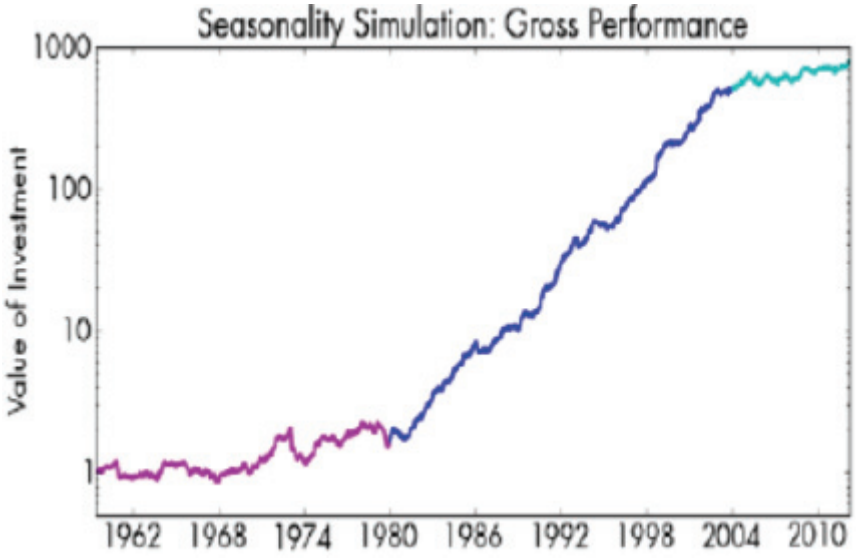
Overfitting occurs when a model has too many parameters relative to the amount of data available to fit the values of those parameters. The consequence is that the model is tuned to noise as well as signal, so it reproduces the data on which it was fitted very well. But if faced with new data generated by the same true underlying process, the model will not fit this new data well.[7] An example of overfitting is shown in Figure 2. Here the dataset was generated with a linear model (with 2 parameters) and some noise, but we have fitted models with up to 10 parameters. The more complicated models will appear to fit better in that they go through (or near) more data points. But, such a model is not expected to predict new data very well if that data comes from the linear model.

Figure 2: An example of overfitting; 30 data points generated from a 1st-order polynomial (two parameters), subsequently fitted with polynomials of order 1, 4, and 9. The higher order polynomial will have a lower average residual between the data and the model fit. But it is unlikely to fit new data.



In finance, this is an easy trap to fall into - where improving the apparent goodness of fit translates into historical simulations that appear to be highly profitable. An example of overfitting is shown in Figure 3 which shows the back-tested performance of a seasonal trading strategy.

Figure 3: An example of a seasonal trading strategy that has been overfitted. The model parameters were chosen based on the period 1980 to 2004. When applied to the out- of-sample periods the performance is significantly worse than during the in-sample period.



The parameters of the strategy were chosen based on the period 1980 to 2004 and the strategy was then applied to the out- of-sample periods before 1980 and after 2004. The performance in both of these periods is significantly worse than the in-sample performance. If only the post-2004 period was tested we might attribute the drop in performance to the seasonal effect being arbitraged away in a more efficient market. However, the similarly poor performance prior to 1980 strongly suggests that overfitting is the culprit.

Overfitting is a subtle but real problem. After all, it's only sensible to choose a model and its parameter values because it does well on past data. It would be rather perverse to deliberately choose a model which did poorly on past data! Clearly a careful balance is needed, and finding a good model involves sophisticated statistical methods.

3.2 Selection Bias

Models with a large number of free parameters are most prone to overfitting. But what if we try hundreds of separate simple models and retain only those that best fit the data? The result is the same as overfitting in the case of multiple parameters, although here there are multiple hypotheses and the resulting effect is usually referred to as selection bias.

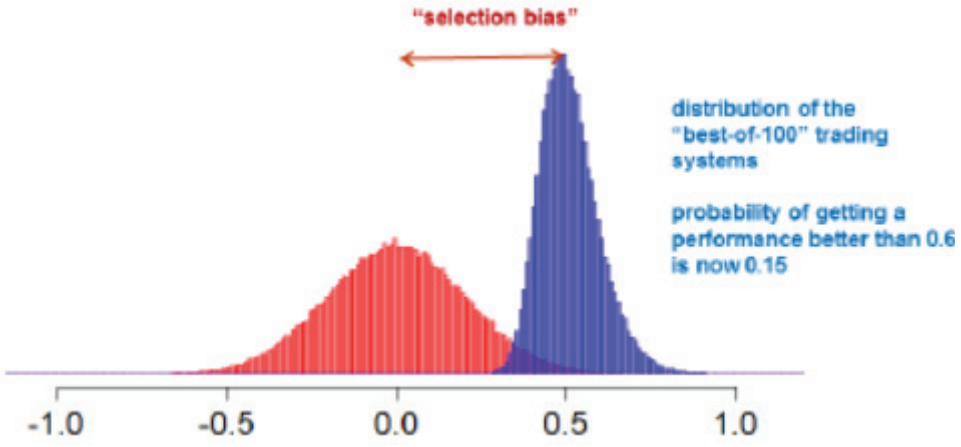
Selection bias can occur when one researcher tests many ideas, but it can also occur more subtly at an institutional level. Imagine that one hundred researchers each test the efficacy

of a single variable for predicting market movements. Each individual researcher might be meticulous about avoiding over fitting and correctly compute their p-values. But, if only the one researcher who found the highest level of predictability reported their results to management then this paints a misleading picture.

As an example, suppose a researcher presents a trading strategy that has a historic Sharpe ratio of 0.6. In Figure 1, the distribution of Sharpe ratios from random, skill-free, strategies over the same time period for this example was calculated. Only 0.2% of these random strategies have Sharpe ratios better than 0.6, meaning we were confident about its predictive power.

However, if this trading strategy was presented because it was actually the best of 100 tested strategies, then it is not appropriate to compare its Sharpe ratio with those in Figure 1. It should be compared with a distribution of the best Sharpe ratios as selected from 100 random strategies, each of which have no real power. This is the blue distribution in Figure 4. In this distribution, 15% of the skill-free strategies have Sharpe ratios exceeding 0.6. In general, a p-value of 0.15 is not deemed to be statistically significant - it is quite probable to achieve such a result without any true predictive power.

Figure 4: The performance distribution of individual skill-free trading strategies, over 23 years is shown in red. The performance distribution of a skill-free trading strategy cherry-picked as the best performer from a set of 100 is shown in blue. In this example, the probability of obtaining a Sharpe ratio better than 0.6 is only 0.002 in the individual distribution but increases to 0.15 in the "best-of-100" distribution.



A memorable lesson in the dangers of multiplicity was given when neuroscientists placed a dead salmon in a magnetic resonance imaging machine. The dead salmon was asked to look at photographs depicting human emotions, and the scientists detected apparently

statistically significant activity in parts of its brain. The problem (as the scientists were trying to illustrate) was that by looking for activity anywhere in the salmon's brain they were effectively testing thousands of hypotheses. When they corrected for this, the result was no longer significant.[8] Again, the problem is by no means just an academic one. In trying to find a good model, a researcher will study many ideas as it would be perverse simply to try one.

3.3 Publication Bias

The selective reporting of results is a recognised issue in many fields of science including medicine, psychology, economics and political science.[9] [10] [11] [12] It might arise because researchers are unmotivated to write-up "negative" results or because scientific journals don't want to publish such results. In areas such as clinical drug testing there might even be commercial incentives to conceal negative findings. Whatever the cause, this selectivity gives rise to publication bias, also known as the file drawer effect in reference to the unpublished negative results hidden away in file drawers. This selection process will mean that a disproportionate number of results published in the scientific literature are nothing more than statistical flukes. Therefore subsequent experiments will be unable to reproduce these results; a drug which appeared to work, loses its efficacy.

There is no reason to think finance is immune to these problems, and many reasons to think it is particularly susceptible. The short-term gains of a positive result will push people to chase products and strategies that they can show "work" on past data. Evidence exists that supports this suspicion, such as poor performance in the mutual fund industry[17], and more recently studies have shown that hypothetical performance data is not consistent with live performance results in the case of ETFs and CTAs.[13] [14] This could be due to malicious mis-selling or poor analysis (not taking into account transaction costs, for example) but publication bias could also be contributing.

3.4 Survivorship Bias

Survivorship bias arises when the performance of a model affects its inclusion in a study. For example, by selecting the largest 10 CTAs, you would be inadvertently selecting funds which have probably performed well in the past. Therefore, a dataset of their past performance will not be representative of the wider industry, or provide reliable forecasts of what we can expect to happen in the future. Survivorship bias is a significant problem in finance, where dead funds often disappear from people's memories and databases.

Consider a set of funds with no skill. Some will produce decent returns simply by chance and these will attract investors, while the poorly performing funds will close and their results may disappear from view. Looking at the results of those surviving funds, you would think that on average they do have some skill. But they don't and in the future they should not be expected to perform as well as they did in the past. Even if the funds really

do have investment skill, the best performers will still tend to be the ones that have also been lucky. Therefore we should not expect the best performers in the population to do quite so well in the future - there will be some "regression to the mean" in their results.

4. SOLUTIONS

The numerous sources of bias might seem rather overwhelming, but there are a number of statistical tools and processes that can help mitigate the problems. Most generally, one should invoke "Occam's razor" which states that the simplest of the possible explanations is the one to be preferred. A more quantitative solution to the issue of overfitting is cross-validation, which applies the principle that the same data used to develop a hypothesis cannot also be used to test that hypothesis.

A non-statistical strategy for tackling the file-drawer problem is to force researchers to publish the results of all their studies. This strategy is being increasingly adopted for clinical drug trials, and many countries now require that trials are pre-registered. The aim is to ensure that the literature is representative of all research, rather than a potentially biased subset of research as is currently the case.[15] [16]

It is beyond the scope of this paper to provide a thorough and technical discussion of suitable methods, for which we refer the reader to.[18]

5. Discussion

The aim of statistical inference, whether it is in trading strategy development, medical research, evaluation of public policy, or any other area is not to say what would have been most effective in the past, but to make a statement about what will hold in the future. For valid inferences, we need to avoid bias in our conclusions.

Bias can be introduced into the scientific literature, as well as into estimates of the performance of trading strategies, in a variety of ways. Individual researchers can introduce bias by overfitting their models or cherry-picking their predictor variables or model structures. Institutions can introduce bias by inadvertently operating policies which tend to favour the reporting of positive results. Universities, academic journals, and financial modellers are all open to the risk.

This paper has described the causes of such bias and some of the tools which can be used to reduce the problem, yielding valid scientific conclusions. Unreliable results can also be the product of deliberate mis-selling, or even incompetence. These are quite different from the problems discussed here, which we believe are subtle enough to fool many smart investors, scientists, and statistical researchers inside the financial industry, as shown by recent studies.[13] [14]

ENDNOTES

1. There is no verifiable source for Mark Twain having said this so ironically it may itself be an example of something that just ain't so.

2. Trouble at the lab, The Economist, 23–27, Oct. 19th–25th, 2013. [2]

3. D. Sarewitz, Beware the creeping cracks of bias, Nature, 485, 149, 2012.

4 .J. P. A. Ioannidis, Why Most Published Research Findings Are False, Public Library of Science: Medicine, 8, 2005.

5. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, Psychological Science, 22, 1359–1366, 2011.

6. The Sharpe ratio measures the risk-adjusted average return. It is the ratio of the mean return and the standard deviation of returns.

7. Data which was not used in the original analysis, but comes from the same underlying process, is referred to as “out-of- sample" data. While data used during the model fitting is called “in-sample".

8. C. M. Bennett, et al., Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiplecomparisons correction, Human Brain Mapping Conference, San Francisco, Jun. 2009.

9. J. P. A. Ioannidis, Contradicted and initially stronger effects in highly cited clinical research, Journal of the American Medical Association, 294, 218–228, 2005.

10. E. J. Masicampo, D. R. Lalande, A peculiar prevalence of p values just below .05, The Quarterly Journal of Experimental Psychology, DOI: 10.1080/17470218.2012.711335, 2012.

11. C. Doucouliagos, T. D. Stanley, Are all economic facts greatly exaggerated? Theory competition and selectivity, Journal of Economic Surveys, 27, 316–339, 2013.

12. A. Gerber, N. Malhotra, Do statistical reporting standards affect what is published? Publication bias in two leading political science journals, Quarterly Journal of Political Science, 3, 313- 326, 2008.

13. J. M. Dickson, S. Padmawar, S. Hammer Joined at the hip: ETF and index development, Van- guard research (2012)

14. M. Beddall, K. Land, The Hypothetical Performance of CTAs, Winton Working Paper (2013)

15. C. Chambers et al., Trust in science would be improved by study pre-registration, The Guardian, Letters, Jun. 5, 2013.

16. E-J. Wagenmakers et al., An agenda for purely confirmatory research, Perspectives on Psycho- logical Science, 7, 632–638, 2012.

17. E. F. Fama, K. R. French, Luck Versus Skill in the Cross Section of Mutual Fund Returns, Journal of Finance, 5, 1915–1947, 2010. [15] D. J. Hand, The Improbability Principle, Scientific American/Farrar, Straus and Giroux, 2014.

18. C. M. Bennett, et al., Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple.

DISCLAIMER

Mark Roulston is Director of Research at Winton Capital Management. David Hand is Senior Research Investigator and Emeritus Professor of Mathematics at Imperial College, London and Chief Scientific Advisor to Winton Capital Management.