WINTON

RESEARCH

Experiment and Observation in Quantitative Investment Management

We assess which methods are most pertinent for different trading strategies, and explain why a robust research process is vital to Winton's approach.

28 OCTOBER 2019 - 6 MINUTE READ

Quantitative investment management resembles the natural sciences, in that it attempts to understand phenomena through empirical analysis of data. It is common for researchers to use techniques developed for scientific analysis to predict market movements and create novel trading strategies.

Experimental vs Observational Science

One aspect of the scientific process that is significant for quantitative investment management is the distinction between *experimental* and *observational* investigations. Experiments can be repeated many times to generate large datasets of comparable results. To take an example from physics: a particle accelerator smashes protons together to see how often a Higgs particle is produced from the resulting debris. If more data are needed to measure more precisely the likelihood of Higgs production, the process can be repeated (budget allowing) until the desired level of precision is reached. This may generate billions or trillions of data points.

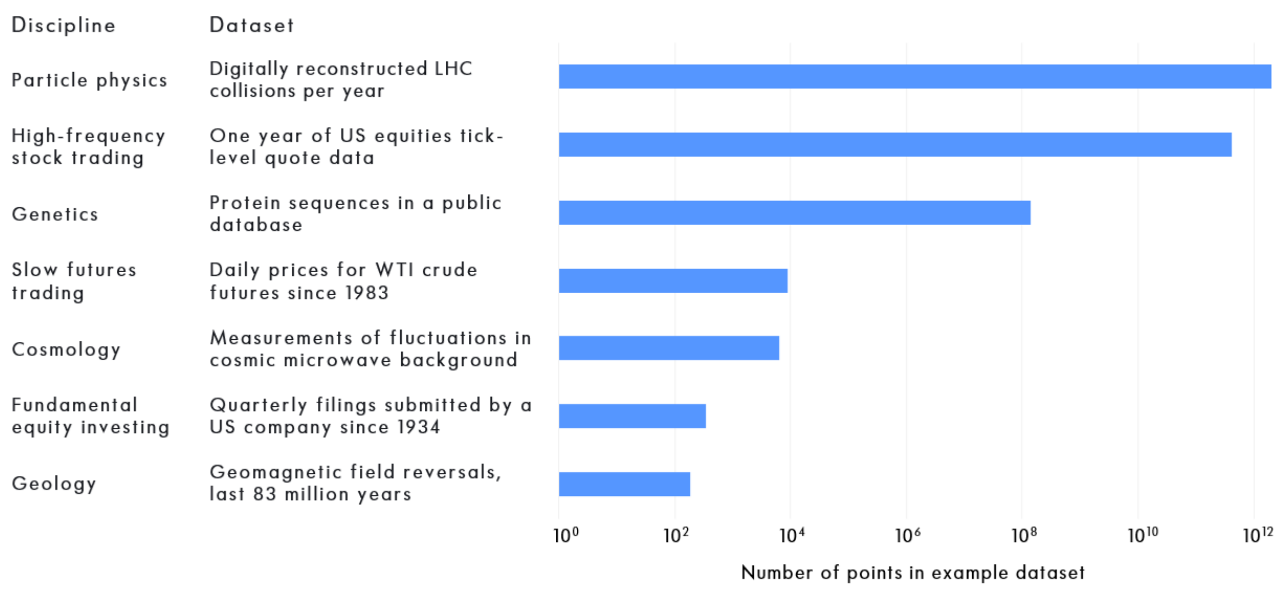
Contrast this experimental approach with an astronomer studying the gravitational waves produced by colliding black holes. The astronomer examines as many actual collisions as possible, and theorises about the precise details of the resulting waves. There is no way to create a large number of comparable pairs of black holes and set them in motion to see what occurs. Moreover, the relative scarcity of black holes makes it likely that the set chosen will not be a representative sample and thus contain biases. In this observational approach, astronomers must take the universe as they find it, try to correct biases in their data, and draw conclusions from the available information.

However, experiment and observation are not a rigid dichotomy but rather two directions of travel on a continuous scale. The more it is possible to control and repeat the process of creating relevant data, the further one can move towards the experimental end of the scale. Going in the other direction, datasets become smaller in size and more prone to various forms of bias, with signals often harder to distinguish from noise.

The distinction between experimental and observational research extends to finance. For example, an execution algorithm can be tested experimentally by applying it to additional trades. Yet a theory about stock-market crashes can be studied only observationally, since the only data comes from crashes that have already taken place, with widely varying circumstances in each case. More crashes cannot be generated on demand!

points, since most listed companies issue financial statements only quarterly. At the other, high-frequency traders process prices stamped to the nanosecond, with the resulting datasets comparable in size to those used in particle physics.

Comparison Between the Sizes of Example Financial and Scientific Datasets



Explanatory notes: The particle physics example shows annual number of collisions at the Large Hadron Collider that pass an initial filter and are digitally reconstructed, assuming it runs for 70% of the year; The genetics example shows number of protein sequences in RefSeq database; The cosmology example shows combined number of multipoles in the TT, TE and EE spectra from “Planck 2018 results. VI. Cosmological parameters”, Planck Collaboration 2018; The geology example shows number of shifts in the polarity of the Earth’s magnetic field, visible in spreading sea floor rocks.

As suggested by the above figure, the different amounts of data available to high-frequency and low-frequency traders determine their positions on the experiment-observation scale. Both types of trader invest in the same markets, but high-frequency traders use the vast amount of data that is available at shorter timescales. This enables them to operate experimentally, because they have more data against which to assess their ideas.

Faster trading signals are more suited to experiments for another, more practical reason: they often have higher expected Sharpe ratios, which means that they can be judged more quickly by their out-of-sample performance.

If, for example, a signal with a supposed Sharpe ratio of 2 or more has produced losses after a few months, it is likely that something is wrong – perhaps the idea has been widely discovered and “arbitraged out”. The signal can then be switched off, and a new experiment started. By contrast, a strategy with an estimated Sharpe ratio of 0.5 could be down for more than a year, but it would not be rational to stop trading the strategy due to performance alone. This is because a multi-year loss is consistent with the long-term expected statistical distribution for that level of Sharpe ratio.

Experiment and observation thus represent two different approaches to quantitative investment. The experimental method involves looking for faster strategies with higher Sharpe ratios. Individually, these strategies will have limited capacity because of the transaction costs incurred by their relatively frequent trades. The aim, however, is to build a large portfolio by combining lots of fast signals.

The alternative approach is to seek out signals that have higher capacity, and generally lower Sharpe ratios. Fewer such signals are needed to build a large portfolio, provided they have low correlation to one another.

As in the scientific arena, experiment versus observation is a spectrum, not a binary choice. It is, nevertheless, a useful framework for understanding the alternative methods used by different quantitative investment managers. We summarise some of the main differences in the following table.

construction	Sharpe ratio, low-capacity signals.	are uncorrelated and have high capacity.
Hypotheses versus data	Data-driven, since the ability to discard loss-making strategies quickly allows for rapid experimentation on new datasets.	Form a hypothesis and then search for the data required to test it.
Explicability	No need to understand why the algorithm chooses to make a particular trade. Any one signal will only have a small allocation and will be switched off quickly if it stops working.	Understanding why the strategy takes its positions is vital. Confidence in a signal can increase if there is a rationale for why it works.
Back test length	Long backtests are unnecessary as signal weights are adjusted rapidly in line with performance.	Longer backtests (ideally 30 to 40 years) help to build confidence in performance and correlation properties of slower signals.
Statistical aims	Extracting information from vast quantities of recent data.	Making reliable inferences when the data available is limited.
Statistical methods	Rich playground for machine learning techniques, which require lots of input data and do not necessarily provide explicable outputs.	Some scope for machine learning, but more focus on simpler mathematical techniques based on a deeper understanding of the relationship between the data and the world.
Selection bias	Less of a concern given the ability to respond to short-term performance.	A major concern. The performance of low Sharpe ratio signals cannot be judged immediately so it is vital to mitigate biases at the research stage.
Scientific inspiration	The application of machine learning and high-performance computing to domains, including image and voice recognition, and natural language processing.	Medicine, gravitational waves, extrasolar planets, etc. Fields that are characterised by low signal-to-noise ratios and a high risk of selection bias.

Although Winton has done more on the experimental side over the past few years, historically our approach has been more observational. Part of the reason stems from our roots in trend following on futures markets, a paradigmatic example of a slow trading signal with large capacity and a relatively low Sharpe ratio. As a result, much of our research has been structured to find other uncorrelated signals to combine with momentum. We focus on the statistical tools and analysis methods best suited to this approach.

Use of Machine Learning

The rapid recent increase in the amount of data available in just about every sphere has created new possibilities for predictive modelling.

For example, a traditional equity analyst might read every report produced by or about the companies they cover, and may in the past have known every relevant fact or figure about a specific company when making an earnings forecast. The data used in an earnings forecast today, however, could include satellite imagery, credit card spending information, logistical details of every product on every truck, and much more besides.

It would be impractical for a person or group of people to pay as close attention as in the past to this massively increased amount of data. But this apparent problem is a great opportunity to apply a collection of statistical techniques known colloquially as “machine learning”. From recognising the content of images to making targeted recommendations on retailers’ websites, the success of these methods has been extraordinary. But they have one vital requirement: a lot of input data.

Hence the applicability of machine learning to faster trading strategies: the volume of short-term price information generates large amounts of data. With slower trading systems, the comparatively limited information content of small, noisy datasets is a less appropriate input for machine-learning models. In such cases, it is more useful to work on drawing reliable conclusions from the data, and to concentrate on interpretability and simplicity rather than applying unnecessarily complex algorithms.

WINTON

perform a lengthy backtest. By way of example, consider a trading strategy that analyses the text of quarterly company reports. To perform a 40-year backtest on the largest 1,000 US companies, we would need to analyse 160,000 reports. And any change to algorithms using the data would require all the reports to be analysed afresh. This task is beyond a group of individuals. Machine learning methods are appropriate instead.

The Danger of Selection Bias

Winton's research is largely hypothesis-driven. A researcher will start with an idea about the world, and then search for data to test whether it is correct. When it comes to creating a trading signal from the idea, the aims are low correlation to existing strategies, low turnover, and a Sharpe ratio that is low but positive, perhaps in the range of 0.3 to 0.5.

These goals may sound modest. Yet finding just 16 uncorrelated signals, each with a Sharpe ratio of 0.5, would result in a portfolio with a Sharpe ratio of 2 and a very large capacity – a far less modest ambition!

This is difficult to achieve in practice, however. While it is easy to create a backtest of a trading signal with a low Sharpe ratio, it is extremely hard to be sure that the Sharpe ratio will remain positive in future. Statistical estimation error is one issue. A much more pernicious problem is selection bias.

To grasp how selection bias operates, imagine coming up with 100 random trading signals, which by definition have no insight or power to predict market movements. Nevertheless, their performance in a backtest will not be precisely zero, but will form a distribution, and some of the signals will seem to have Sharpe ratios of 0.3 or higher. If we pick only those with positive historical performance, while discarding the others, we will have created a portfolio with an attractive backtest that in reality has a Sharpe ratio of zero, or less, after including transaction costs.

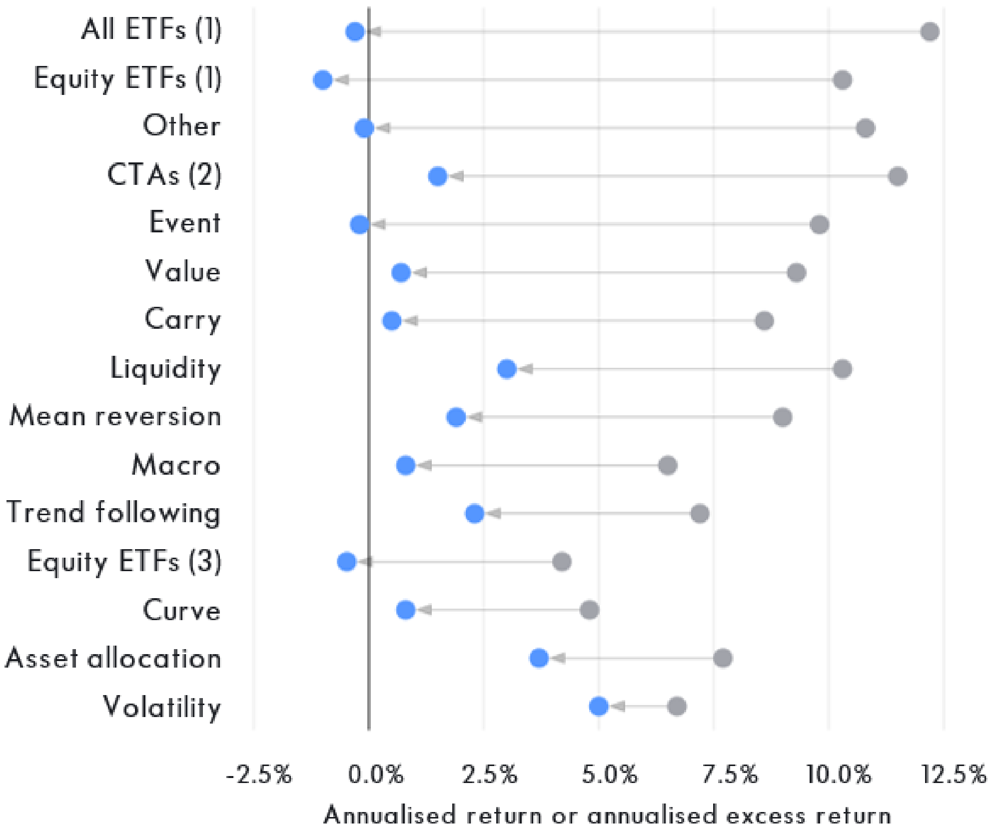
This is a stylised example, but unfortunately the reality of research can be all too similar. Researchers test lots of ideas, and although they are not generated randomly, we do not know in advance whether they work or not. Even if the ideas are good on average, the best backtests will be partly the result of the idea working, and partly luck, so the true Sharpe ratio is likely to be overestimated.

Moreover, the selection bias arising from picking the ideas with the best backtests is deeply ingrained in many organisational cultures. Employees want to show their managers only their best results. When something works less well, it is easy to file it away and move on to an idea that looks more promising, or to adjust the parameters of the model until it does work. Even when researchers are aware this is happening, they often fashion subsequent explanations for why the original idea would have failed, and so exclude it from the list of ideas they have tried.

A seminal 2005 paper with the dramatic title "Why Most Published Research Findings Are False" widely publicised the effect of selection bias in academia, where the selection is often at the point of publication. Journals are more likely to print papers that report significant results than those that do not. This has resulted in the so-called "replication crisis", where researchers have been unable to reproduce the results of earlier work.

A parallel in investment management is the disparity between the backtest and live performance. We have shown previously the tendency for trend-following products to underperform their backtests after launch. A meta-analysis drawing data from various sources shows that the problem appears across the investment landscape.

Their Launch



Strategies are ordered by size of performance degradation. In-sample performance is shown by the grey dot and out-of-sample performance, by the blue dot. The results are those published in A. Suhonen, M. Lennkh and F. Perez, Quantifying Backtest Overfitting in Alternative Beta Strategies, Winter 2017, except for: (1) Vanguard, Joined at the Hip: ETF and Index Development, July 2012; (2) Winton, Hypothetical Performance of CTAs, December 2013; (3) C. Brightman, F. Li and X. Liu, Chasing Performance with ETFs, November 2015.

Mitigating Selection Bias

As indicated above, the problem of selection bias is not purely technical. It can appear even if all researchers carry out their work to an exemplary standard. The problem is rather with the framework under which the research is organised. The necessity of addressing the issue at an organisational level is discussed in a recent paper in the context of machine learning applications in quantitative finance. Here we describe some of the steps our more observational research process entails.

The key structure at Winton for mitigating selection bias is the idea of a hypothesis register. This takes inspiration from the idea of the clinical trials register, which helps to reduce the effect of selection bias in medical research. At Winton, a proposed new signal is recorded in precise detail in the register, which is then visible to the whole research department. This gives other researchers the opportunity to peer review the idea at an early stage. Importantly, it also allows us to keep track of the number of ideas that we test. The collaborative and open nature of the process also helps to reduce the pressure on individual researchers to share only positive findings.

Just as in the case of a clinical trial, the registry includes precise details of how the idea will be tested: what data will be used, what time period will be used for different parts of the analysis, what statistical tests will be performed, and so on.

It is important that the registration of a trading idea includes every small variation that will be tested. As we have seen, if too many ideas are tested then there is a higher probability that a spurious success will appear by chance. A higher significance threshold is needed if we test more variations on an idea, which means we need to know at the very least how many ideas we are testing.

The idea is tested on market data once this process is completed. The result is a much more robust research framework. We can reject trading strategies that would otherwise be accepted, had we failed to keep track of the number of related ideas being tested and not performed the necessary statistical corrections. And we are able to produce a much more accurate assessment of the likely out-of-sample performance of our signals.

WINTON

Quantitative investment managers operate across a spectrum. One end can broadly be characterised as closer to trading, usually involving higher-frequency strategies; a more experimental approach to implementing new systems; a focus on higher Sharpe ratios and lower capacities; a requirement for large (often intraday) datasets, and an attendant interest in machine learning. The other end is closer to investment, with generally slower systems with lower transaction costs; an approach that is necessarily more observational; sub-strategies with lower Sharpe ratios but capable of managing more capital; and a requisite appreciation of the subtleties of dealing with small amounts of data and finding weak signals in noisy datasets.

Building confidence in lower Sharpe ratio strategies, or in a process to research them, is difficult. The difficulty is as much organisational as technical and requires a significant top-down approach, analogous to the way many governments now mandate the registration of medical trials. At Winton, for example, every hypothesis tested by our researchers since 2012 has been pre-registered. We have also run meta-experiments over several years on our research methodology to test its effectiveness, with success.

Amid the fervent hype, big data and machine learning do offer opportunities for investment managers pursuing slower trading strategies. But more often researchers face the problem of making inferences from relatively small amounts of data. In such cases, it is critical to use the techniques discussed to extract reliable information.

RESEARCH PRACTICES SYSTEMATIC INVESTING

Related Articles

Economies of Scales: How to Chart Long-Term Time Series

Charts are widely used in finance. As well as representing information efficiently, they can reveal relationships or anomalies in data. But their effectiveness or otherwise is a function of the choices made during chart construction.

Using Differential Privacy to Protect Personal Data

Systematic Methods for Classifying Equities

Our research explores alternative approaches for assigning sectors to the S&P 500 using stock price covariance and natural language processing.

Winton is a research-based investment management company with a singular focus on statistical and mathematical inference. Founded in 1997, Winton manages money for many of the world's largest investors.

Follow us

Subscribe

Subscribe to the latest updates from Winton.

Your email

Privacy
Policies and Disclosures
Terms of Use