UNIGESTION

# DESIGNING A ROBUST BACKTESTING FRAMEWORK

December 2018

Joshua Seager, Ravi Ramakrishnan

# Table of Contents

# ABSTRACT

Academic research has shown that backtested results are often far superior to those in live performance tests. There are logical explanations for this. Multiple testing increases the probability of stumbling across strategies that look significant but lack economic value. When one does find a strategy with economic value, overfitting is both tempting and easy. We have designed a backtesting process that aims to minimise the probability of making these mistakes and realising disappointing returns when strategies go live.

The process guards against overfitting primarily through the backtest register, which requires that all backtests are registered prior to the start of testing. We also run a robustness check on the final strategy to ensure that the strategy does not vary too much with small, economically irrelevant changes to parameters. This would be a sign of an over-fitted strategy which can lead to disappointing live performance.

Another key consideration is the statistical significance of results. This can be hard to establish, especially when one considers that backtested results are often the product of multiple tests. We use a bootstrap process within a multiple hypothesis-testing framework to tackle this issue. We build distributions of simulated 'best' Sharpe ratios under the null hypothesis. These help us infer the significance of our selected strategy within the context of the number of strategy variants we analysed in the backtesting process.

# INTRODUCTION

Backtesting is the process by which we determine how a given strategy would have behaved historically. This is a necessary and important part of researching a new strategy. Using a backtest, we can assess the statistical significance of metrics we observe and build expectations of how the strategy might behave in the future. While this is very useful, the nature and volume of past investment research can mean that backtested results are not as easy to interpret as one might hope.

# PITFALLS

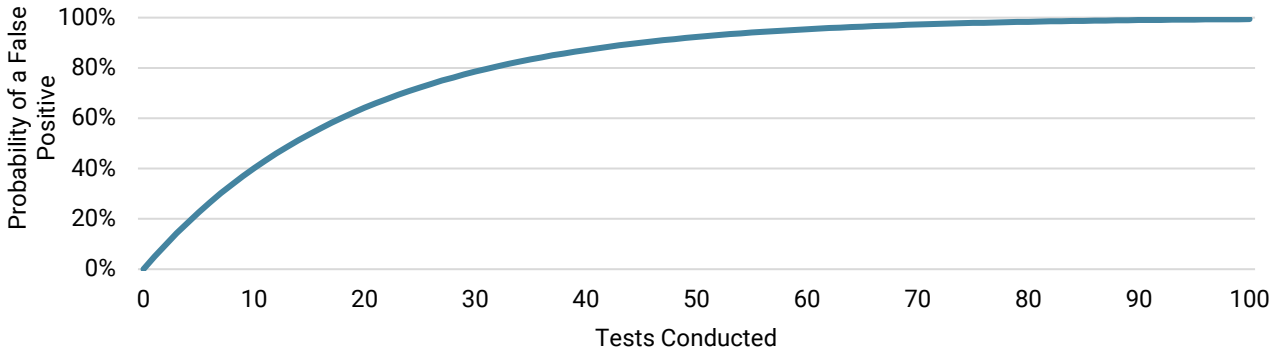
Many of the difficulties inherent in interpreting the results of a backtest derive from the fact that investors and researchers have been building and parameterising strategies on the same historical dataset for many years. This creates a phenomenon called data snooping whereby researchers' conclusions are, in some part, a function of their knowledge of the underlying data. When such a process is in place, there is always the possibility that the success of the developed strategy is attributable to luck rather than economic merit (Sullivan et al, 1999). In such a scenario, we would not expect the success of the strategy to continue into the future.

Suhonen et al (2016) investigate just how frequently this occurs. They compare the backtested and live performance of 215 alternative beta strategies across 5 asset classes and 11 strategy types provided by 15 Investment Banks. The deterioration in risk-adjusted performance is startling. In the backtest, the median Sharpe ratio is 1.2 and 95% of strategies exhibit a positive and significant Sharpe at the 10% level. Once the strategies go live, the median Sharpe falls to 0.3 and just 25% of Sharpes are positive and significant to 10%.

We offer two potential explanations for such a drop in performance. We will refer to these as bias 1 and bias 2.

Bias 1: Over the last 50 years, hundreds of papers have been published detailing backtested investment strategies. Given that many strategies will have been tested and discarded, it's not unreasonable to assume that thousands of strategies have been tested. As the number of strategies tested increases, so does the probability of discovering false positives. In this case, a false positive is a strategy that performs well by luck, not because it has any intrinsic economic value. An analogy is the simultaneous flipping of n coins. If n is sufficiently large, it's likely that we will find a coin that lands on heads each time we flip. Clearly, this does not mean that we have found a special coin and we should not expect this to continue into the future (out-of-sample) (White, 2000). As such, increasing the number of tests conducted without increasing the threshold for statistical significance increases the probability of discovering a false positive (Harvey et al, 2014). While this bias is recognised and tackled in other statistical fields, it is frequently neglected in finance and economics (Fabozzi, López De Prado, 2018).

**Chart 1:** Likeihood of false positive at 5% significance level increases with number of tests




Source: Unigestion,

Bias 2: In a perfect world, researchers would use one set of data to build and parameterise strategies, and a completely different set to test them. Unfortunately, however, there is only one history of asset prices. Even if a researcher diligently sets aside a portion of data for out-of-sample testing, it's possible that this data was used to produce analysis that could have guided the researcher towards the strategy they are testing. As such, it's very hard to find truly out-of-sample data and many strategies are effectively built and tested on the same dataset. This increases the probability of overfitting, where a strategy is too finely tuned to the data that it was devised and tested on (Sullivan et al, 1999).

# POTENTIAL SOLUTIONS

The simplest way to deal with these biases is by applying a haircut to backtested Sharpe ratios: industry practice is around 50% (Harvey, Liu, 2015). This is a crude approach, which fails to account for the different ways that strategies are developed. For example, is it sensible that a strategy derived from economic theory using one backtest receives the same haircut as a data-driven strategy, which is the result of hundreds of backtests? Probably not (Harvey et al, 2014). Furthermore, empirical evidence suggests that a blanket 50% haircut is not appropriate. Harvey and Liu (2015) find that the appropriate haircut for multiple testing is actually non-linear. Strategies with very high Sharpes are likely true discoveries; therefore, the 50% haircut is too harsh. Conversely, once you have accounted for multiple testing, strategies with marginal Sharpes tend to be revealed as false positives, which require discarding.

A more nuanced option is to use a multiple backtesting framework. This implicitly accounts for bias 1 by measuring the statistical significance of a result in the context of the number of tests conducted to reach said result. There are a number of ways to do this, which allow one to control for either the probability of one false positive or the overall proportion of false positives. The simplest is the Bonferroni method where the p-value of a strategy is multiplied by the number of tests that were considered in the multiple hypothesis test. This is a very conservative method, which works well if the number of tests (N) is small but becomes extremely stringent as N increases. For example, if N = 1000, a pre-adjustment p-value of 0.005% is required for a significant result at a 5% confidence level. There are other, less stringent methods, like Holms adjustment, and Benjamini, Hochberg and Yekutieli's adjustment, which work by ordering tests based on p-value and adjusting the threshold for significance sequentially.

There are, however, difficulties with these approaches. Most obviously, one needs to pick an N, which represents how many other hypothesis tests are incorporated into the assessment of significance. Many multiple testing methods also become excessively stringent when test statistics are dependent (Harvey et al, 2014). Using an extreme example, when testing 1000 perfectly correlated strategies, there is no need to diverge from single hypothesis testing as the test is really on just one strategy. Applying Bonferroni's adjustment to this set of strategies would simply increase the threshold for significance by a factor of 1000.

Another option is to build a distribution of possible test statistics for a given strategy using simulations. One can then compare the test statistic of the chosen strategy to the simulated test statistics to assess significance. This is particularly useful for strategies that have non-normal properties because it does not require any assumptions about the distribution of the strategy returns.

# OUR APPROACH

There are four key elements of our backtesting approach which we believe protect us against the biases described above. To illustrate our approach, we will use an equity value strategy that we developed in 2018. The strategy is relatively simple: going long equity markets deemed cheap and short markets deemed expensive through futures.

## 1. Backtest Register

Prior to backtesting a strategy, we define the variants that we will test in a backtest register. This method is used in fields outside of finance, notably in medical testing, to encourage a rigorous testing process. The International Committee of Medical Journal Editors (ICMJE) employ a similar procedure. To be considered for publication, clinical trials must be registered in a public trials registry at, or before, the time of first patient enrolment.

Prior to developing the equity value strategy, we selected three main variations that we wanted to test. These are listed below.

a. Using unprocessed valuation metrics for the trading signal. Five tests were run at this stage using two allocation methodologies.

b. Using normalised (5 year z-score) valuation metrics for the trading signal. We tested three different normalised metrics and a combination of all three using a simple, rank based allocation.

c. Normalised metrics for the trading signal and more advanced allocation methods. Each test used the combined z-score (decided at the end of stage 2 due to its superiority) as the signal whilst varying the weighting methodology or risk adjustment in the allocation method.

**Table 1: Equity Value Backtest Register**

| | Backtest Group | Signal | Allocation Method | Annual Return | Volatility | Sharpe Ratio | Max Drawdown |
|---|---|---|---|---|---|---|---|
| 1 | Raw Signal | Earn Yield | Distance from median | 1.64% | 7.06% | 0.23 | 26.88% |
| 2 | | Roll Yield | Distance from median | -3.96% | 7.90% | -0.50 | 54.44% |
| 3 | | Div Yield | Distance from median | -1.28% | 7.38% | -0.17 | 39.92% |
| 4 | | Earn Yield | Risk adjusted distance from median, beta neutral | 1.20% | 6.48% | 0.19 | 25.89% |
| 5 | | Dividend Yield | Risk adjusted distance from median, beta neutral | -0.08% | 6.50% | -0.01 | 29.03% |
| 6 | Normalised Signal | Earn Yield–Z Score | Rank | 3.05% | 6.92% | 0.44 | 11.93% |
| 7 | | EV/EBITDA–Z Score | Rank | 3.00% | 6.71% | 0.45 | 13.89% |
| 8 | | Div Yield–Z Score | Rank | 2.53% | 7.23% | 0.35 | 13.50% |
| 9 | | Combined–Z Score | Rank | 3.01% | 5.84% | 0.52 | 10.44% |
| 10 | Allocation Variations | Combined–Z Score | Risk adjusted rank | 3.01% | 6.08% | 0.50 | 12.12% |
| 11 | | Combined–Z Score | Distance from median | 3.56% | 6.39% | 0.56 | 10.19% |
| 12 | | Combined–Z Score | Distance from median, beta neutral | 4.11% | 6.17% | 0.67 | 8.10% |
| `13 | | Combined–Z Score | Distance from median, beta neutral | 4.13% | 6.17% | 0.67 | 10.32% |

Source: Unigestion, Bloomberg. Data for the period 1999-2018

The backtest register serves two purposes. Firstly, predefining the variants that we will test focuses our attention on adjusting parts of the strategy that our intuition suggests should be important as opposed to those that optimise backtested performance. For example, when we developed the equity value strategy, it would have been tempting to test a huge range of lookback periods and valuation metrics. It is likely that this would have improved backtested performance. However, there is

no reason why a seven-year lookback should significantly outperform a five-year lookback, or that an obscure valuation metric should be a superior signal than a combination of three well-known metrics. As such, there is no reason to assume that the performance gained from making these changes will persist into the future. It is more probable that such changes simply increase the likelihood of building an over-fitted strategy and, in turn, realising disappointing live performance. The backtest register removes this temptation entirely, thereby guarding against bias 2. Secondly, as table 1 demonstrates, it allows us to log and report the number of tests conducted and how each test has performed. This prevents selective reporting of results and shows how the strategy returns have evolved as we have made changes to the process.

## 2. Tail Risk, Regime Conditional and Scenario Analysis

As with all other parts of our investment process, we use broad measures of risk such as expected shortfall and maximum drawdown during backtesting. These measures incorporate asymmetry, kurtosis and illiquidity into the estimated risk of a strategy.

We analyse how the strategy would have performed in our four market regimes: Growth, Inflation, Recession and Market Stress. This develops our understanding of when the strategy will perform but also acts as a sense check. Prior to developing a strategy, we have a view of the risks that it is exposed to and the environments that it should perform well in. If our analysis of regime-conditional performance contradicts this intuition, it suggests that we do not understand the strategy, or that we have not constructed it effectively. Both instances would present cause for concern and require further analysis. Table 2 shows a summary of this analysis for the equity value strategy, which performs well in periods of steady growth but suffers in market stress. This is in line with academic research, which suggests that value risk premia compensate the investor for distress risk. As such, this analysis confirms that the strategy behaves as anticipated.

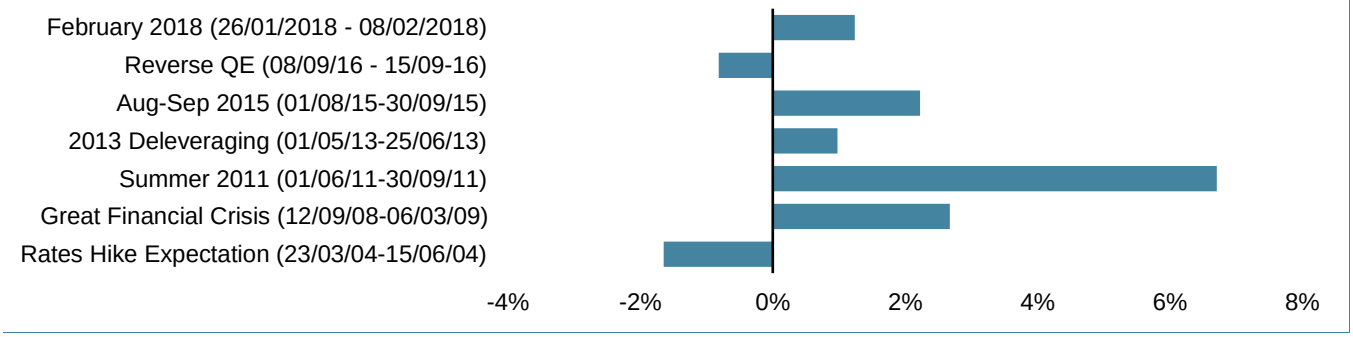
**Table 2: Regime Conditional Returns**

| | Recession | Inflation Shock | Market Stress | Steady Growth | Full Period |
|---|---|---|---|---|---|
| Return | 3.96% | 3.47% | -0.33% | 5.55% | 4.29% |
| Sharpe Ratio | 0.72 | 0.63 | -0.06 | 1.01 | 0.78 |
| Sharpe (T-Stat) | 1.17 | 1.04 | -0.08 | 2.57 | 2.92 |
| Hit Ratio | 0.55 | 0.56 | 0.45 | 0.61 | 0.57 |
| Skewness | 0.60 | 0.44 | 0.14 | 0.18 | 0.21 |
| Kurtosis | 2.89 | 2.51 | 2.97 | 2.73 | 2.96 |

Source: Unigestion, Bloomberg. Monthly data for the period 1999–2018

Where possible, we also conduct scenario analysis. We analyse how the strategy performed in periods of historical market stress and, where possible, hypothetical scenarios. For the latter analysis, we first regress the strategy against relevant variables, either generic macro or more specific to the strategy. We then stress these variables and measure the performance impact. This allows us to see how the strategy would perform in hypothetical environments. Chart 2 shows the historical scenario analysis for the equity value strategy, which confirms that the strategy has not exhibited any unexpected tail behaviour historically.

**Chart 2: Historical Scenario Analysis**



Source: Unigestion, Bloomberg. Data for the period 1999–2018

## 3. Bootstrap

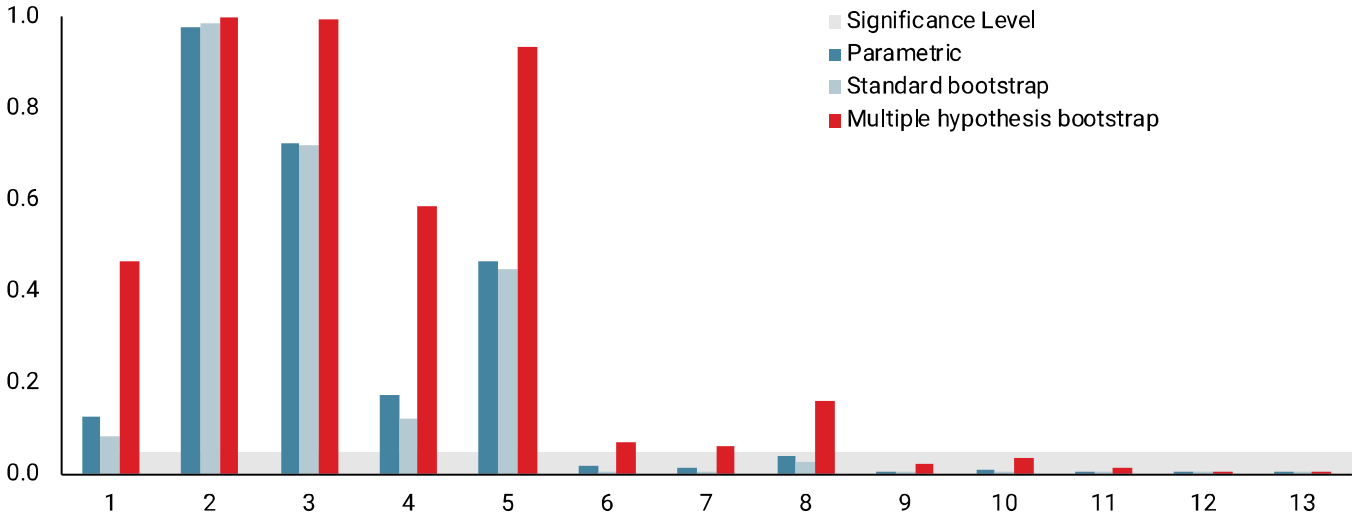
We use a bootstrap process to assess the statistical significance of our results. Given a dataset, bootstrapping refers to building a new dataset by resampling from the original dataset with replacement. Statistics of interest are computed on each resampled dataset to make inferences about the original data. For example, one may estimate the variance of a statistic by calculating it many times on resampled datasets. This variance can be used to infer the statistical significance of the original statistic. Within the backtesting framework, we use bootstrapping to build a distribution of Sharpe ratios under the null hypothesis (that our strategy really has an expected return of zero). We then check where the Sharpe ratio of our strategy sits on this distribution to infer the probability that it too comes from the null distribution. We use two methods for this.

Standard bootstrap: We bootstrap the demeaned returns of the selected strategy to generate time series from the null hypothesis H0. A block bootstrap method is used to preserve autocorrelation of returns. We then compute the Sharpe ratio of each bootstrapped dataset to build a distribution of H0 Sharpe ratios. The selected strategy's Sharpe ratio is compared to this distribution to compute the bootstrap p-value. The p-value allows us to infer the statistical significance of the Sharpe ratio of our strategy. Intuitively, the p-value here shows how likely it is that a strategy with the same shaped distribution as ours, but no economic value, would produce the Sharpe that we observed in our backtest purely by luck.

Multiple hypothesis bootstrap: While the method above is useful, if we have run 1000 backtests to find a strategy with a high return and low volatility, it is likely to continue to look significant. This is the multiple hypothesis testing effect (bias 1), which we tackle using the approach of White (2000). Instead of bootstrapping the demeaned returns of only the selected strategy, we bootstrap the demeaned returns of all strategies from the backtest register. We compute the Sharpe ratios for each bootstrapped sample and select the maximum Sharpe ratio per bootstrap. The maximum Sharpe ratio from each bootstrap is added to the distribution of "best" Sharpes. The p-value is calculated by comparing the Sharpe ratio of the selected strategy to the distribution of best Sharpes. By construction, this will give us a more conservative p-value estimate. Intuitively, taking the maximum Sharpe at each bootstrap simulates the effect of backtesting multiple strategies and picking the best. This needs to be considered because, as demonstrated earlier, the likelihood of a false positive increases as we test more strategies. As such, the p-value here shows how likely it is that we would have found a Sharpe as high as we did from a strategy with no economic value given the number of strategies we tested. We favour this method because it very clearly focuses on the multiple testing risks of bias 1.

We will compare the results of both bootstrap analyses as well as the more standard parametric calculation (Mertens, 2000) on the 13 backtests that we ran when developing the equity value strategy (details in table 2). To conduct the standard bootstrap analysis, we first demeaned the returns of each strategy. We bootstrapped these, building 13 distributions of simulated Sharpe ratios, one for each strategy. The p-values were calculated by comparing the Sharpe of a strategy with the distribution of simulated Sharpes. For the multiple hypothesis bootstrap, we also used the record of the weekly returns of each strategy from the backtest register. Again, we subtracted the mean return from each variant time series to set the expected strategy value to 0. These returns were used to generate 10,000 simulations of each variant (13 x 10,000 in total) under the null hypothesis. For each of the 10,000 simulations, we calculated all 13 Sharpe ratios and selected the maximum. After each simulation, we added the maximum of the 13 Sharpe ratios to the distribution of 10,000 'best' Sharpe ratios. We then calculated p-values of all 13 strategies by comparing the Sharpe ratio of each strategy to this distribution of 'best' Sharpes. We also used a parametric method for comparison. The results of all methods are displayed in chart 3 below.

**Chart 3: Strategy P-Values**



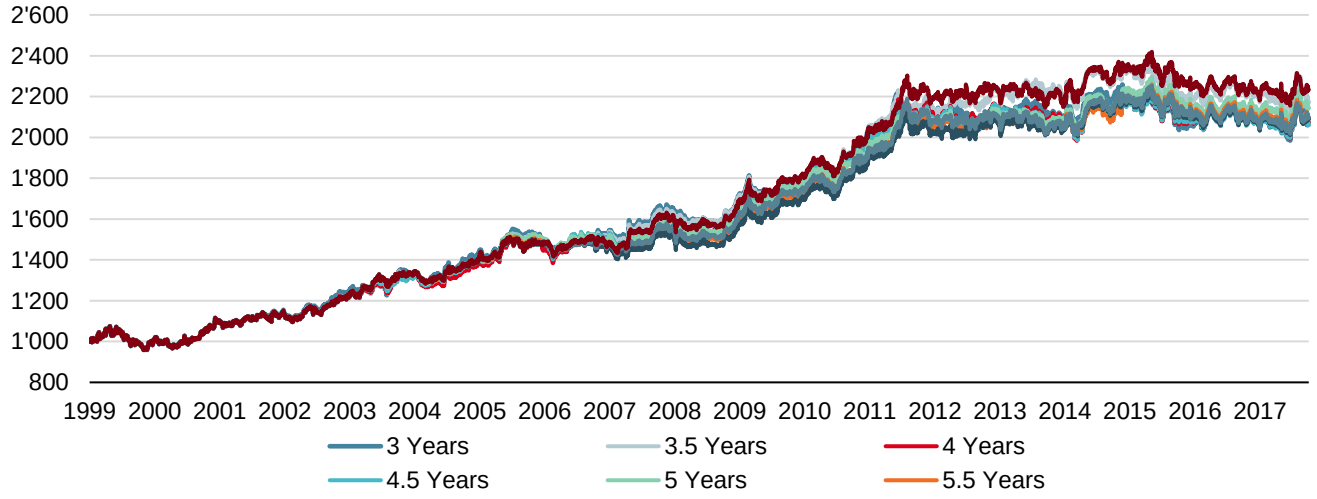

Source: Unigestion, Bloomberg. Data for the period 1999–2018

What is important to note is that strategies 6, 7 and 8 look significant at the 5% level when considered as single tests (parametric and standard bootstrap) but not once we account for the number of tests conducted (multiple hypothesis bootstrap). Practically, this means that, had we discovered them using only one test, we would have accepted those Sharpe ratios as statistically significant. However, once we account for the fact that we conducted 13 tests, those Sharpe ratios fall within the normal statistical variance that we would expect from 13 strategies from the null hypothesis. This shows that when one fails to account for multiple hypothesis testing, it's easy to mistake insignificant results for significant ones. In an investment context, this exposes you to disappointing live performance. The strategy that we picked, number 12, shows robustness to all three tests, which gives confidence that our strategy has economic value.

## 4. Robustness Check On Final Strategy

Finally, we ran a robustness test on the selected strategy by adjusting key parameters and rerunning the strategy. Examples of such parameters would be the lookback period used to calculate a signal, the way signals are blended or even the universe of tradeable instruments. For the equity value strategy, we conducted a number of these tests, adjusting the investment universe and also the lookback period used to normalise the valuation metrics. Chart 4 below shows the results of the lookback-based robustness check.

**Chart 4: Strategy NAV Using Different Lookback Periods**



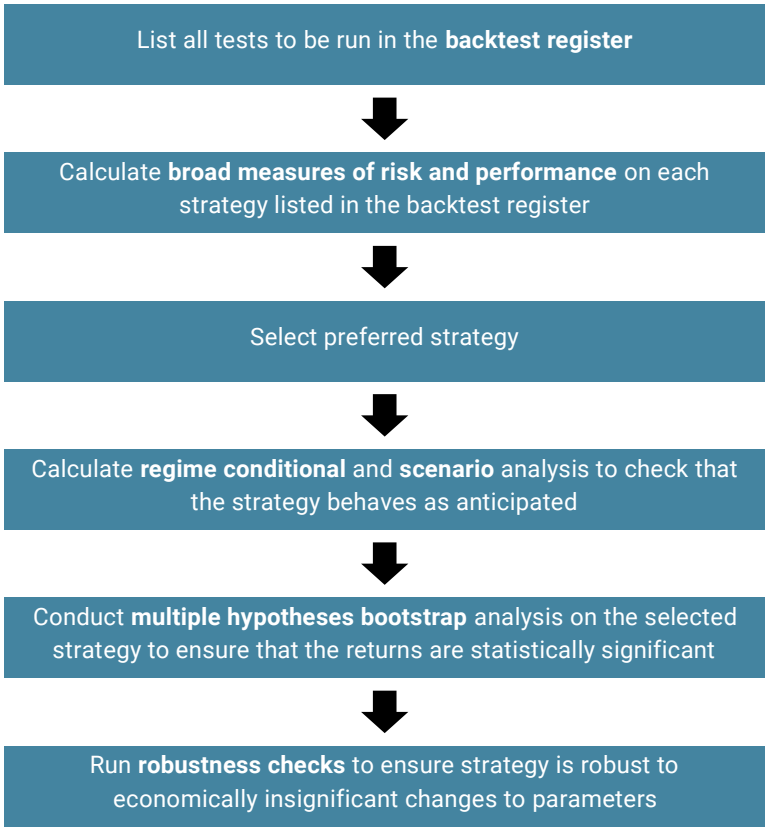

Source: Unigestion, Bloomberg. Data for the period 1999–2018

The aim here is to ensure that the final strategy is robust to economically insignificant changes in input variables. The stability of the performance of the equity value strategy demonstrates that this is the case. This is important because if there

are good economic reasons for a value strategy to be successful, it shouldn't make a huge difference whether we measure value relative to the last three, five or seven years. If the strategy behaves completely differently as we make changes that should be economically irrelevant, there is a chance that the economic rationale for the strategy is not valid or the strategy is overly fitted to the data.

Therefore, strong robustness to these changes guards against both bias 1 and bias 2. More practically, we need to pick a set of parameters to use. While we can pick the best historical parameterisation, we have no way of knowing the best parameterisation for the future. While this isn't linked to a particular backtesting bias, high sensitivity to parameterisation adds another layer of variability to our expected returns, which we would rather not have.

The flowchart below summarises the full backtest process that each strategy goes through if it is to be traded live in a Unigestion portfolio. We believe that this mitigates the common backtesting biases described earlier in the document as much as possible.




For illustrative purposes only

# REFERENCES

F. Fabozzi, M.López De Prado (2018) – "Being Honest in Backtest reporting : a template for disclosing multiple tests", Journal of Portfolio Management, Forthcoming

C.Harvey, Y.Liu, and H. Zhu (2014) – "…and the Cross-Section of Expected Returns", Review of Financial Studies, 29, 5-68.

C.Harvey, Y.Liu (2015) – "Backtesting" - Journal of Portoflio Management 42, 13-28

R.Kosowski, N.Naik, M.Teo (2007) – "Do Hedge Funds Deliver Alpha? A Bayesian and Bootstrap Analysis", Journal of Financial Economics, 84, 229-264

H.R. Kunsch (1989) – "The Jackknife and the Bootstrap for General Stationary Observations", The Annals of Statistics, Vol. 17, No. 3, 1217-1241

R.Y.Liu, K. Singh (1992) – "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence.", In: R. Lepage and L. Billard, Eds., Exploring the Limits of Bootstrap, John Wiley, New York

E.Mertens (2002) – "Comments on variance of the IID estimator in Lo", Research Note, http://www.elmarmertens.org/.

D.N. Politis, H. White (2004) – "Automatic Block-Length Selection for the Dependent Bootstrap" Econometric Reviews Vol. 23, No. 1, 53–70

R. Sullivan, A.Timmerman, H.White (1999) – "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap" – Journal of Finance 54, 1647-1691

A.Suhonen, M.Lenkh, F.Perez (2016) – "Quantify Backtest Overfitting in Alternative Beta Strategies " Journal of Portfolio Management 43, 90-104

H. White (2000) – "A Reality check for Data Snooping " Econometrica 68, 1097-1126

http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

# APPENDIX

## 5. Bootstrap Details

An important issue with bootstrapping, especially when dealing with time-series data, is that of correlation. In a typical vector time-series dataset, we encounter autocorrelation as well as cross-correlation. We deal with cross-correlation by simply bootstrapping the entire vector series as one, as opposed to doing as many bootstraps as there are series. But there is no clear way of dealing with autocorrelation. Here, our preferred approach is the moving-block bootstrap method of Kuensch (1989) and Liu and Singh (1992). In this method, given a vector time series, we specify a block length, choose random starting indices and construct blocks of resampled data from the original data starting with these indices. We then place these blocks next each other in a chained fashion to form the new multivariate bootstrapped time series. Note that though the bootstrap starting indices are computed based on the multivariate data, these indices form a vector so that within the bootstrapped data, the original cross-correlation is maintained. Then there is the question of the block length. Here again, we use the Politis and White (2001) method to estimate the optimal block length. Assuming the block length follows a geometric distribution, this estimate is derived by minimising the fourth moment of this distribution.

# DISCLAIMER

## Non-Financial Promotion disclaimer:

This document has been prepared for your information only and must not be distributed, published, reproduced or disclosed by recipients to any other person.

This is not a financial promotional. The document It constitutes neither investment advice nor recommendation. This document represents no offer, solicitation or suggestion of suitability to subscribe in the investment vehicles it refers to. Please contact your professional adviser or consultant before making an investment decision.

Some of the investment strategies described or alluded to herein may be construed as high risk and not readily realisable investments, which may experience substantial and sudden losses including total loss of investment. These are not suitable for all types of investors. To the extent that this report contains statements about the future, such statements are forward-looking and subject to a number of risks and uncertainties, including, but not limited to, the impact of competitive products, market acceptance risks and other risks. As such, forward looking statements should not be relied upon for future returns.

Data and graphical information herein are for information only and may have been derived from third party sources. Unigestion takes reasonable steps to verify, but does not guarantee, the accuracy and completeness of this information. As a result, no representation or warranty, expressed or implied, is or will be made by Unigestion in this respect and no responsibility or liability is or will be accepted. All information provided here is subject to change without notice. It should only be considered current as of the date of publication without regard to the date on which you may access the information. Rates of exchange may cause the value of investments to go up or down. An investment with Unigestion, like all investments, contains risks, including total loss for the investor.

## Back testing disclaimer:

Back-tested or simulated performance is not an indicator of future actual results. The results reflect performance of a strategy not currently offered to any investor and do not represent returns that any investor actually attained. Backtested results are calculated by the retroactive application of a model constructed on the basis of historical data and based on assumptions integral to the model which may or may not be testable and are subject to losses.

Changes in these assumptions may have a material impact on the backtested returns presented. Certain assumptions have been made for modeling purposes and are unlikely to be realized. No representations and warranties are made as to the reasonableness of the assumptions. This information is provided for illustrative purposes only. Backtested performance is developed with the benefit of hindsight and has inherent limitations. Specifically, backtested results do not reflect actual trading or the effect of material economic and market factors on the decision-making process. Since trades have not actually been executed, results may have under- or over-compensated for the impact, if any, of certain market factors, such as lack of liquidity, and may not reflect the impact that certain economic or market factors may have had on the decision-making process. Further, backtesting allows the security selection methodology to be adjusted until past returns are maximized. Actual performance may differ significantly from backtested performance.