Literature Review

1 Backtesting and Model Validation

Much of financial academic literature is currently facing a problem in terms of validation and verification of results. The primary method of going about these ends in the past has been to perform historical simulations, or backtests , in order to prove profitability of a trading strategy. The recent advances in both technology and the algorithms available to construct these strategies has resulted in researchers being able to run so many iterations of a model or strategy configuration through these backtests, that its become increasingly difficult to control for spurious results, with some papers suggesting that most published research findings are false [9].

The standard way of implementing backtests is to split the data into two portions: an In Sample (IS) portion which is used to train the model, and an Out of Sample (OOS) portion which is used to test the model and validate results. The problem lies in that millions of different model configuration might be tested, and if more sophisticated test measures are not in place (i.e. just the standard Neyman-Pearson hypothesis testing framework is implemented), then it is only a matter of time before a false positive result occurs which shows high performance both IS and OOS (i.e. overfitting). The nature of financial data, where there is a low signal-to-noise ratio in a dynamic and adaptive system, and where there is only one true data sequence, makes it difficult to resolve these issues effectively [2][11].

Overfitting is not a novel issue, and has of course been tackled in various literature areas, including machine learning. However, in that context, the frameworks are often not suited to the buy/sell with random frequency structure of investment strategies. They also do not account for overfitting outside of the output parameters, or take into consideration the number of trials attempted. Other methods, such as hold-out, are arguably still faulty due to research knowledge while constructing models [13]. One of the downfalls of the typical IS-OOS set up in the financial context, is also that the most recent (and relevant) data will not be able to be used for the model training.

There have been some suggestions to resolve the problem that is occuring in the literature as a result of this - some work suggesting new frameworks, which this section will cover, and others which focus on the review process or how data and replication procedures are made available. While the points made with regard to the review process and so on are certainly important, they dont aid with more effective model training for the researcher up front, and so will not be covered here [12].

1.1 Testing Methodologies

Considering the issues laid out above, there has been much work to develop alternative approaches to backtesting. One of the common approaches to avoid backtest overfitting is the hold-out strategy, where a certain portion of the dataset is reserved for testing true OOS performance. Numerous problems have been pointed out with this approach, including that the data is often used regardless, or that awareness of the movements in the data may, consciously or otherwise, influence strategy and test design by the researchers [13]. For small samples, a hold-out strategy may be too short to be conclusive [14], and even for large samples it results in the most recent data (which is arguably the most pertinent) not being used for model selection [8][2].

There has been work by several authors to try and lay out techniques to try and avert backtest overfitting. The Model Confidence Set (MCS), as developed by Hansen et al. [5], starts with a collection of models or configurations, and remove models iteratively according to a defined loss function. The confidence set is defined by the remaining models once a non-rejection takes place within the process, and these models are considered to be statistically similar within a certain confidence range. MCS is thus able to facilitate equitable model selection. However, Aparicio et al. [1], showed that while MCS is a potential strategy, in practice is is ineffective due to the inordinate requirement of signal-to-noise necessary to identify true superior models, as well as a lack of penalization over the number of trials attempted.

Bailey et a. [2] have developed a more robust approach to backtesting and how overfitting during strategy selection might be avoided. Their research defines backtest overfitting as having occurred when the strategy selection which maximizes IS performance systematically underperforms median OOS in comparison to the remaining configurations. They

use this definition to develop a framework which measures the probability of such an event occuring, where the sample space is the combined pairs of IS and OOS performance of the available configurations. The probability of backtest overfitting (PBO) is then established as the likelihood of a configuration underperforms the median IS while outperforming IS.

Formulaically, the definition of backtest overfitting is given by

$$\sum_{n=1}^{N} E[\overline{r_n}|r \in \Omega_n^*]Prob[r \in \Omega_n^*] \leq N/2 \tag{1}$$

Where the search space Ω consists of the N ranked strategies, and their IS performance r and OOS performance $\bar{r}$ This allows the PBO, using the bayesian formula, to be defined as

$$PBO = \sum_{n=1}^{N} Prob[\overline{r} < N/2|r \in \Omega_n^*]Prob[r \in \Omega_n^*] \tag{2}$$

Notably, the above definitions considers IS as the data made available to the strategy selection, rather than the models calibration (e.g. the full IS dataset, rather than, by was of example, the number of days used in a moving average). This allows the model-free and non-parametric nature of the definition.

They further developed the combinatorially symmetric cross-validation (CSCV) framework as a methodology to reliably estimate the probability used in PBO, which allows a concrete application of the concept. The CSCV framework does not require using the typical hold-out strategy (and thus avoids credibility issues), and is ultimately able to provide a bootstrapped distribution of OOS performance.

The methodology can be briefly summarised (skipping some details and nuances) as the following steps:

1 Generate a TxN performance series matrix, M, representing the profits and losses by the N trials over T time periods

2 Partition the M matrix into S submatrices

3 Generate the combination set C of all combinations of the S submatrices

4 For each combination in C:

   a Form the training set by joining the 2 combination sets, and testing set as the rest of the combinations (all in order)

   b Determine the ranked in-order IS and OOS performance for the sets

   c Determine n* as the best performing IS strategy

   d Determine the relative rank of the n* strategys OOS performance, where we should observe that $\bar{r}^*$ systematically outperforms OOS as well Define logit $\lambda = \frac{\overline{w}_c}{(1-\overline{w}_c)}$, where a high value implies consistency between IS and OOS performance, and thus a low level of backtest overfitting

5 The λvalues can then be collected and used to define the relative frequency at which λoccurs across all combination sets in C, signified by $f(\lambda)$.

The CSCV framework and results thus allows the consideration of several notable statistics. First and foremost, the PBO may now be estimated using the CSCV method and using an integral over the $f(\lambda)$ function as defined above which offers a rate at which the best IS strategies underperform the median of OOS trials. If $\varphi \approx 0$, it is evidence of no significant overfitting (inversely, $\varphi \approx 1$ would be a sign of probable overfitting). Critically then, a PBO measure may be used in a standard hypothesis test to determine if a model should be rejected or not. This can be extended, as shown by Bailey et al., to show the relationship between overfitting and performance degradation of a strategy. It becomes clear that with models overfitting to backtest data noise, there comes a point where seeking increased IS performance is detrimental to the goal of improving OOS performance.

The CSCV methodology provides several important benefits (some already mentioned) over traditional testing frameworks, including the usual K-fold cross validation used in machine learning. By recombining the slices of available data, both the training and testing sets are of equal size, which is particularly advantageous when comparing financial statistics such as the Sharpe Ratio (SR), which are susceptible to sample size. Additionally, the symmetry of the set combinations in CSCV ensure that performance degradation is only as a result of overfitting, and not arbitrary differences in data sets. It is pointed out that while CSCV and PBO should be used to evaluate the quality of a strategy, they should not be the function on which strategy selection relies, which in itself would result in overfitting.

1.2 Test Data Length

The CSCV methodology offers an important but highly generalised framework to assess models and backtest overfitting. It doesnt however indicate which metrics should be used to assess the IS and OOS performance, nor any indication on the amount of data needed to do so effectively. One of the noted limitations of the framework is that a high PBO indicates overfitting within the group of N strategies, which is not necessarily indicative that none of the strategies are skillful - it could be that all of them are. Also, as pointed out, it should not be used as an objective function to avoid overfitting, but rather as an evaluation tool. To this end it helps assess overfitting, but not necessarily avoid it.

A typical measure of evaluation used for financial models is the Sharpe Ratio (SR), which is the ratio of between average excess returns and the returns standard deviation - a measure of the return on risk. In the context of comparing models, SR is typically expressed annually to allow models with different frequencies to be compared [10] shows

$$SR = \frac{\mu}{\sigma}\sqrt{q} \tag{3}$$

Using sample means and deviations, $\hat{\mu}$ and $\hat{\sigma}$, SR can be shown to converge as follows (as y $\to \infty$)

$$\hat{SR} \to \mathcal{N}[SR, \frac{1 + \frac{SR^2}{2q}}{y}] \tag{4}$$

Thus, when using SR estimations, which follow a Normal distribution, it is possible that where the true SR mean is zero we may still (with enough configurations attempted) find an SR measurement which optimises IS performance. This is shown by Bailey et al. [**?**], who propose the non-null probability of selecting an IS strategy with null expected performance OOS. Notably, typical methods such as hold-out once again fail, as the number of configurations attempted are not recorded. They add a further derivation, thich is the Minimum Backtest Length (MinBTL), ultimately showing that

$$MinBTL \approx (\frac{(1-\gamma)Z^{-1}[1-\frac{1}{N}] + \gamma Z^{-1}[1-\frac{1}{N}e^{-1}]}{\overline{E[max_N]}})^2 < \frac{2ln[N]}{\overline{E[max_N]}}^2 \tag{5}$$

The statistic highlights the relationships between: selecting a strategy with a higher IS SR than expected OOS, the number of strategies tested (N), and the number of years tested (y). The equation shows that as the number of strategies tested increases, the minimum back test length much also increase in order to contain the likelihood of overfitting to IS SR.

As shown extensively throughout ML literature, increased model complexity and number of parameters is one of the primary causes of overfitting. In context of the MinBTL formula, model complexity affects the number of configurations that are available and which may be tested, which in turn will increase likelihood of overfitting. A lack of consideration, or reporting, of the number of trials makes the potential for overfitting impossible to assess.

Bailey et al. expanded on this view with assessing the impact of presenting overfit models as correct. T hey were able to show that in lieu of any compensation effects (i.e. a series following a Gaussian random walk), there is no reason for overfitting to result in negative performance. However, where compensation effects apply (e.g. economic/investment cycles, bubble bursts, major corrections etc.), then the inclusion of memory in a strategy is likely to be detrimental to OOS performance if overfitting isnt controlled for [3].

1.3 Sharpe Ratio

The use of the Sharpe Ratio in financial backtesting is not just an arbitrary or persistent literature choice. The statistic offers two benefits: the effectively strategy-agnostic financial information contained, as well as being relatable to the t-statistic, and so simple to perform hypothesis testing. The SR ratio (estimate from sample as $\hat{SR}$) is defined as

$$SR = \frac{\mu}{\sigma} \tag{6}$$

The t-ratio is defined as

$$t - ratio = \frac{\hat{\mu}}{\hat{\sigma}/\sqrt{T}} \tag{7}$$

Evidently, the link here is trivial, as per formula (3) . As noted earlier though, the chances of overfitting with the SR ratio, even if true mean returns are zero, are relatively significant. Once of the strategies employed to try and counteract this is to use a haircut, where the reported SR ratio is discounted by 50

The 50that in a context of multiple testing, the haircut is nonlinear - the highest Sharpe ratios are moderately penalized, whereas the marginal Sharpe ratios were heavily penalized. While initially fairly sensible, Harvey et al raise valid concerns regarding the effect on option strategies, controlling for risk as well, pertinently, what constitutes an appropriate level of significance testing. In light of this, they develop a p-value based statistic for multiple testing, the haircut adjusted sharpe ratio HSR, as well as expand upon work by [7] to provide a distribution that can be used in a dependent multiple testing framework with an appropriate p-value adjustment.

This work is relevant, in that the HSR statistics proposed offer another framework in which investment strategies might be evaluated against each other. The primary difference in comparison to PBO and CSCV, is that where they offer a methodology for evaluating strategies within a group, HSR aims to adjust the statistical significance of each strategy such that the overall risk of spurious correlation is controlled for. A benefit of this method is that there is less chance of a relevant strategy being disregarded as a result of just poor peer performance. PBO however, does have the primary benefit of being metric-agnostic, where the HSR framework is largely based on using the Sharpe ratio (though it can be generalized to another statistic with a probabilistic interpretation). Additionally, PBO is generally more in line with machine learning literature with the cross validation like approach on time series data.

It should be noted, that the literature detailing usage of the Sharpe ratio for strategy comparison is extensive, with numerous variations and methodologies offered [4]. However, the crux of this paper lies in whether an online neural network is able to make effective enough predictions that a strategy might use the predictions to be profitable. The subtlety here is that we will consider the usage of such forecasting within a strategy, rather than as a strategy itself. In line with this, statistics such as the Sharpe ratio will be used, but not form a critical consideration of the research here as the comparison of strategies used will be a secondary consideration.

2 References

[1] Aparicio, Diego and Lopez de Prado, Marcos, How Hard Is It to Pick the Right Model? (December 2017). Available at SSRN: https://ssrn.com/abstract=3044740 or http://dx.doi.org/10.2139/ssrn.3044740

[2] Bailey, David H. and Borwein, Jonathan and Lopez de Prado, Marcos and Zhu, Qiji Jim, The Probability of Backtest Overfitting (February 27, 2015). Journal of Computational Finance (Risk Journals), 2015, Forthcoming. Available at SSRN: https://ssrn.com/abstract=2326253 or http://dx.doi.org/10.2139/ssrn.2326253

[3] Bailey, David H. and Borwein, Jonathan and Lopez de Prado, Marcos and Zhu, Qiji Jim, Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance (April 1, 2014). Notices of the American Mathematical Society, 61(5), May 2014, pp.458-471. Available at SSRN: https://ssrn.com/abstract=2308659 or http://dx.doi.org/10.2139/ssrn.2308659

[4] Bailey, David H. and Lopez de Prado, Marcos, The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality (July 31, 2014). Journal of Portfolio Management, 40 (5), pp. 94-107. 2014 (40th Anniversary Special Issue).. Available at SSRN: https://ssrn.com/abstract=2460551 or http://dx.doi.org/10.2139/ssrn.2460551

[5] Hansen, Peter Reinhard and Lunde, Asger and Nason, James M., The Model Confidence Set (March 18, 2010). Available at SSRN: https://ssrn.com/abstract=522382 or http://dx.doi.org/10.2139/ssrn.522382

[6] Harvey, Campbell R. and Liu, Yan, Backtesting (July 28, 2015). Available at SSRN: https://ssrn.com/abstract=2345489 or http://dx.doi.org/10.2139/ssrn.2345489

[7] Campbell R. Harvey Yan Liu Heqing Zhu, 2016. " and the Cross-Section of Expected Returns," Review of Financial Studies, vol 29(1), pages 5-68.

[8] Hawkins, Douglas. (2004). The Problem of Overfitting. Journal of chemical information and computer sciences. 44. 1-12. 10.1021/ci0342472.

[9] Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

[10] Lo, Andrew W., The Statistics of Sharpe Ratios. Financial Analysts Journal, Vol. 58, No. 4, July/August 2002. Available at SSRN: https://ssrn.com/abstract=377260

[11] McLean, R. David and Pontiff, Jeffrey, Does Academic Research Destroy Stock Return Predictability? (January 7, 2015). Journal of Finance, Forthcoming. Available at SSRN: https://ssrn.com/abstract=2156623 or http://dx.doi.org/10.2139/ssrn.2156623

[12] Lopez de Prado, Marcos, The Future of Empirical Finance (May 31, 2015). Journal of Portfolio Management, 41(4). Summer 2015. Forthcoming.. Available at SSRN: https://ssrn.com/abstract=2609734 or http://dx.doi.org/10.2139/ssrn.2609734

[13] Schorfheide, Frank, and Kenneth I. Wolpin. 2012. "On the Use of Holdout Samples for Model Selection." American Economic Review, 102 (3): 477-81.

[14] Weiss, S. M, Kulikowski, C. A. (1991). Computer systems that learn : classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Mateo (Calif.): Kaufmann.