PREDICTING AND PREVENTING OVERFITTING OF FINANCIAL MODELS

AKSHAY CHALANA

This paper is dedicated to Prof. Jim Morrow.

Abstract. Among statistically-driven models, one of the greatest challenges is the prevention of overfitting due to excessive complexity or parameters too specific to the sample set. Given the uniquely chaotic structure of the stock market and the signals into which its progression can be decomposed, common solutions to this do not necessarily apply. We explore a framework outlined by a group of researchers from the Lawrence Berkeley National Laboratory to estimate the probability of backtest overfitting (PBO) from combinatorially symmetric cross-validation and establish a minimum backtest length (MinBTL) in order to effectively fit a model, but avoid overfitting.

Contents

1. Introduction

With the growing prevalence of technical analysis as a basis for amateur trading, the combination of the attraction of extensive testing and tuning of models against historical asset pricing data and increased public awareness of pseudo-mathematical analysis techniques, such as stochastic oscillators, Fibonacci ratios, Elliot waves, and more lead the unaware down the tempting trap of overfitting. Overfitting, with regards to mathematical models, refers to tuning that, either per complexity

Date: May 18, 2017.
Key words and phrases. Backtesting, overfitting, finance, machine learning.

or through standard tuning, accurately predict performance within the sample set, but not outside of it. In a pair of papers, "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance" and "The Probability of Backtest Overfitting," one in *Notices of the American Mathematical Society* and the other in the *Journal of Computational Finance (Risk Journals)*, respectively, David H. Bailey, Jonathan M. Borwein, Marcos Lopez de Prado, and Qiji Jim Zhu explain a modification of *k-fold cross-validation* (K-FCV) and *leave-one-out cross-validation* (LOOCV) known as combinatorially symmetric cross-validation (CSCV) which best caters to time series data, in order to build a distribution of PBOs (probabilities of backtest overfitting) for different model configurations to optimize for the most accurate algorithm, weight for PBO. Beyond building this distribution, this algorithm further evaluates performance degradation, based on metrics such as the Sharpe Ratio, the Sortino Ratio, Jensen's Alpha, and the Probabilistic Sharpe Ratio [2]. The authors further discuss the mathematical premise of a metric referred to as MinBTL, a minimum length of backtests as a function of the number of trials attempting to determine overfitting [3]. In all, though these metrics and algorithms do not provide ideal targets for model optimization so much as evaluations that should instruct decision making post-model generation with regards to investment and advertisement.

2. Definitions

We define a number of the important terms around the problem that is posed.

Definition 2.1. (Backtest Overfitting) Overfitting is characterized by an investment strategy with optimal in-sample performance achieving below-median expected ranking out-of-sample. Application of Bayes' Theorem, this is given by:

$$\sum_{n=1}^{N} E[\overline{r_n}|r \in \Omega_n^*]Prob[r \in \Omega_n^*] \leq N/2 \tag{2.1}$$

where $\Omega_n^* = \{f \in \Omega | f_n = N\}$ and Ω is the the ranking space of "$N!$ permutations of (1,2,...,N) indicating the ranking of the N stratgies" [2].

We also provide the authors' definition of the Probability of Backtest Overfitting.

Definition 2.2. (Probability of Backtest Overfitting) This probability is that of the occurrence above: that a strategy with optimal IS performance receives a below-median ranking OOS.

$$PBO = \sum_{n=1}^{N} Prob[\overline{r_n} < N/2 | r \in \Omega_n^*]Prob[r \in \Omega_n^*] \tag{2.2}$$

Definition 2.3. (Sharpe Ratio) The Sharpe Ratio (SR) quantifies a "strategy's performance on the basis of a sample of past returns" [3]. Per [5], this "is defined as the ratio of the excess exepected return to the standard deviation of return."

$$SR = \frac{\mu - R_f}{\sigma} \tag{2.3}$$

[3] annualizes this as follows:

$$SR = \frac{\mu}{\sigma}\sqrt{q} \tag{2.4}$$

"where q is the number of returns per year" [2]. Note that we know the standard deviation of this distribution to be $y^{-1/2}$. The authors also provide an estimator for the annualized Sharpe ratio:

$$\hat{SR} \xrightarrow{a} \mathcal{N}\left[SR, \frac{1+\frac{SR^2}{2q}}{y}\right] \tag{2.5}$$

Since the SR cannot perfectly be determined for a dataset by a single point or discrete set of points, we refer to its estimate at a point as $\widehat{SR}$.

Though the SR is the primary performance metric referenced in this work, a number of other metrics are referenced, which we shall briefly define.

Definition 2.4. (Sortino Ratio) The Sortino Ratio is a variant of the SR, with distinction made between harmful and total overall volatility through usage of the considered asset's downside deviation, i.e. the standard deviation of its negative returns. It is calculated by subtracting the risk-free rate (rate of return from an investment with no risk; a basis of comparison for additional risk taken in an investment) from the asset's return and dividing by downside deviation. If $< R >$ is the Expected Return, R_f is the risk-free rate of return, and σ_d is the downside deviation, this formula is thus

$$\frac{< R > - R_f}{\sigma_d} \tag{2.6}$$

Definition 2.5. (Jensen's Alpha) Alpha (A_j) measures performance of some asset or portfolio relative to expected return, adjusted for risk. It is calculated using R_p for expected portfolio return, R_f for risk-free rate, B_p for the beta of the portfolio/asset (relative volatility/systematic risk of an asset to the market), and R_m for expected market return:

$$A_j = R_p - [R_f + B_p * (R_m - R_f)] \tag{2.7}$$

Definition 2.6. (Non-Normal $\widehat{SR}$) If we assume returns to not necessarily be sampled from a normal distribution, [4] cites a conclusion by E. Mertens that the SR still follows a normal distribution as follows:

$$(\widehat{SR} - SR) \xrightarrow{a} N\left(0, \frac{1+\frac{1}{2}SR^2 - \gamma_3 SR + \frac{\gamma_4 - 3}{4}SR^2}{n}\right) \tag{2.8}$$

Definition 2.7. (Confidence Band Around $\widehat{SR}$) To account for skewness (a measure of asymmetry of a probability distribution) and kurtosis (the sharpness of the peak of a frequency-distribution curve), we pose a confidence band around $\widehat{SR}$. We manipulate equation (2.8) to determine an estimate on the standard deviation:

$$\hat{\sigma}_{\widehat{SR}} = \sqrt{\frac{1 - \gamma_3 \widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}{n-1}}$$

The $n-1$ here is given by Bessel's correction, a method of correcting for bias. Given significance level α, the confidence band around the true SR is:

$$Prob[SR \in (\widehat{SR} - Z_{\alpha/2}\hat{\sigma}_{\widehat{SR}}, \widehat{SR} + Z_{\alpha/2}\hat{\sigma}_{\widehat{SR}})] = 1 - \alpha \tag{2.9}$$

Definition 2.8. (Probabilistic Sharpe Ratio) Per [4], the Probabilistic Sharpe Ratio (PSR) is an "uncertainty-adjusted investment skill metric," with the purpose of "[correcting]... inflationary effects." We define PSR in terms of a benchmark (can be set by default to 0) SR, SR^*, and an observed SR, $\widehat{SR}$:

$$\widehat{PSR}(SR^*) = Prob[\widehat{SR} > SR^*] = 1 - \int_{-\infty}^{SR^*} Prob(\widehat{SR}) \cdot d\widehat{SR} \tag{2.10}$$

The *cdf* of this distribution is thus

$$\widehat{PSR}(SR^*) = Z\left[\frac{(\widehat{SR} - SR^*)\sqrt{n-1}}{\sqrt{1 - \hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}\right] \tag{2.11}$$

Returning to [2], there are a series of lemmas that the authors rely on to demonstrate the theorem that we shall later discuss.

Lemma 2.9. *We show that* $E[\max_N] = E[\max x_n]$ *for large* N *where* $x_n \sim Z$ *where* Z *is the CDF of the Standard Normal distribution is approximated by:*

$$(1-\gamma)Z^{-1}\left[1 - \frac{1}{N}\right] + \gamma Z^{-1}\left[1 - \frac{1}{N}e^{-1}\right]$$

where $\gamma \approx 0.5772156649$*, the Euler-Mascheroni constant, and* $N \gg 1$*. It can be shown that the upper bound of* $E[\max_N]$ *is* $\sqrt{2\ln[N]}$

Proof. Given that the independent random variables which are sampled follow exponential distributions, their maximum converges asymptotically to a Gumbel distribution, a probability distribution which models the maximum of a set of samples from various other distributions. Its cumulative distribution function is $e^{-e^{-(x-\mu)/\beta}}$. Similarly, per the Gumbel distribution's covering of the Maximum Domain of Attraction of the Gaussian distribution, it also estimates the expected value of the maximum of multiple independent random Gaussian variables. We show this by applying the Fisher-Tippett-Gnedenko theorem to the Gaussian distribution with $G[x] = e^{-e^{-x}}$ as the CDF of the Standard Gumbel distribution and $\alpha = Z^{-1}\left[1 - \frac{1}{N}\right], \beta = Z^{-1}\left[1 - \frac{1}{N}e^{-1}\right] - \alpha$, with Z^{-1} the inverse of the Standard Normal's CDF:

$$\lim_{N\to\infty} Prob\left[\frac{\max_N - \alpha}{\beta} \leq x\right] = G[x]$$

The limit of these maxima normalized, per the Gumbel Maximum Domain of Attraction, is

$$\lim_{N\to\infty} E\left[\frac{\max_N - \alpha}{\beta}\right] = \gamma$$

with γ once again being the Euler-Mascheroni constant. This provides the result originally stated in the Lemma. □

We later provide a definition of cross-validation in outlining the specific algorithm provided by the authors.

3. Minimum Backtest Length

Having established $E[\max_N]$, it is possible to approximate and bound MinBTL. Bailey, et al. state the following:

Theorem 3.1. *MinBTL is set to the strategy out of N with IS Sharpe ratio $\overline{E[\max_N]}$, but expected OOS Sharpe ratio 0:*

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

On one hand, this is stated to imply that this value increases with more configurations, but we also note that it is simply a necessary, not a sufficient condition for overfitting to occur. Bailey, et al. present the example of only having 5 years worth of historical data against which to backtest, which limits the number of model configurations to 45, almost guaranteeing that the model with overfit.

Assuming the distributions used to calculate MinBTL have not been pre-prepared, it is possible to utilize the Combinatorially Symmetric Cross-Validation algorithm mentioned above to determine the PBO (and other useful metrics) by manual testing.

4. Combinatorially Symmetric Cross-Validation (CSCV)

Before outlining the algorithm itself, we briefly define cross-validation. We will later describe the differences betweeen CSCV and the traditional cross-validation methods mentioned above: K-FCV and LOOCV. Cross-validation refers, quite simply, to splitting the initial dataset into training and test sets in order to evaluate the predictive capacity of a model. K-FCV refers to using k partitions of the data, and in k iterative instances using each partition as the validation set, while using the other $k-1$ partitions as training data. A final estimation can be produced through some sort of combination (typically an average) of the outputs of all k separately trained models.

Now, the CSCV algorithm proceeds per the following steps:

1. $T \times N$ matrix M consists of N columns representing N trials, each with T observations of profit or loss per the Nth strategy's performance at that moment.
2. The optimal strategy can eventually be selected by subsampling each column, calculating Sharpe ratio, and choosing the strategy with the highest such ratio.
3. Divide M into S disjoint submatrices.
4. Create combinations of the components of each S of size $S/2$ by the following formula:

$$\binom{S}{S/2} = \binom{S-1}{S/2-1}\frac{S}{S/2} = ... = \prod_{i=0}^{S/2-1}\frac{S-i}{S/2-i}$$

5. From each combination, build training and testing sets and evaluate necessary metrics. This is as follows:
   (a) Join the $S/2$ submatrices in the chosen combination to form a $T/2 \times N$ training set.

(b) Form a testing set $\overline{J} = M \backslash J$ of dimension $T/2 \times N$. These orders are irrelevant for the Sharpe ratio, but are important for other performance metrics, such as return maximum drawdown ratio.
(c) Form order N vector R^c of the performance metric for each strategy (each column of J).
(d) Form a similar vector for $\overline{J}$.
(e) Choose the optimal strategy from the training set based on its ranking in the sorted performance vector.
(f) Determine the relative rank of this strategy by dividing its rank in the testing set by $N + 1$. If overfitting is not occurring, this strategy should outperform this testing ranking OOS.
(g) We determine the *logit* in order to indicate consistency between IS and OOS performance, as a metric for backtest overfitting:

$$\lambda_c = \ln \frac{\overline{\omega}_c}{(1 - \overline{\omega}_c}$$

where ω_c is the aforementioned relative ranking.

(6) Finally, we determine the distribution of ranks by using X as the characterization function and $\#(C_S)$ as the cardinality of C_S:

$$f(\lambda) = \sum_{c \in C_S} \frac{X_{\{\lambda\}}(\lambda_C)}{\#(C_S)}$$

5. Statistics Derived from CSCV

Aside from the aforementioned PBO statistic produced by CSCV, we also discuss the production of 3 other useful metrics: performance degradation of the model over time OOS, probability of monetary loss OOS of the optimal IS model, and stochastic dominance: whether using this strategy of model selection produces preferable results to random selection of model configuration.

5.1. **PBO.** Having performed CSCV, we can now estimate the PBO with the following formula:

$$\phi = \int_{-\infty}^{0} f(\lambda) d\lambda$$

A simple interpretation of this value is as the underperformance of an optimal IS model OOS as compared to the median of all considered models. The ideal case in which this optimal model outperforms the median for the majority of the N trials is signified in the output of the CSCV algorithm by $\lambda_C > 0$. On the other hand, $\phi \approx 0$ indicates a low proportion of outperformance of the median by the optimal model, which, in turn, indicates a low chance of overfitting. By contrast, $\phi \approx 1$ indicates a high likelihood of overfitting. Bailey, et al. pose a set of applications for the PBO:

(1) As an application of the Neyman-Pearson framework, a common statistical tool for considering significance, an investor could simply fight overfitting by rejecting any model determined to be optimal with a PBO determined to be greater than 0.05.
(2) In some Bayesian application, it may be possible to utilize the PBO as a prior probability in determining the posterior probability of a model's prediction.

(3) The PBO can be utilized as a weighting factor for portfolios. Weights could be determined as $(1 - PBO), 1/PBO$ or some other method.

5.2. **Performance Degradation and Probability of Loss.** To discuss the direct negative implications of overfitting, we consider the optimal pair of performance metrics, typically the Sharpe Ratio, but also possible the Sortino ratio, Jensen's Alpha, or the Probabilistic Sharpe Ratio, determined for each combination of models in step 5 of the CSCV algorithm: $(R_{n^*}, \overline{R_{n^*}})$. Since there is no necessary correlation between each of the groups from which R_{n^*} and $\overline{R_{n^*}}$ are taken, even though we know that $R_{n^*} = \max\{\mathbf{R}\}$, it is possible that $\overline{R_{n^*}} < \max\{\overline{\mathbf{R}}\}$. Thus, per compensation effects such as "overcrowded investment opportunities, major corrections, economic cycles, reversal of financial flows, structural breaks, bubble bursts, etc." [3], regressing $\overline{R_{n^*}}^c = \alpha + \beta R^c_{n^*} + \epsilon^c$ practically results in a negative β. Beyond regression, another statistic, the proportion of combinations with negative performance, $Prob[\overline{R_{n^*}}^C < 0]$, can be used to indicate the occurrence of poor performance not actually driven by overfitting, but rather by other factors. This and other statistics can be derived by plotting the aforementioned pairs: $(R_{n^*}, \overline{R_{n^*}})$.

5.3. **Stochastic Dominance.** We seek a definition for stochastic dominance which demonstrates that the OOS performance of an optimally selected strategy is superior to one which is randomly selected. In the first-order, we define this as the case in which

$$\forall x : Prob[\overline{R_{n^*}} \geq x] \geq Prob[Mean(\overline{\mathbf{R}}) \geq x]$$
$$\exists x : Prob[\overline{R_{n^*}} \geq x] > Prob[Mean(\overline{\mathbf{R}}) \geq x]$$

Visually, this is indicated by the cdf of $\overline{R_{n^*}}$ being at or below that of $\mathbf{R}$.

Aside from first-order stochastic dominance, Bailey et al. also present second-order stochastic dominance, which is cited as a "less demanding criterion" [2]. The formula for this requires the following:

$$\forall x : SD2[x] = \int_{-\infty}^{x} (Prob[Mean(\overline{\mathbf{R}}) \leq x] - Prob[\overline{R_{n^*}} \leq x])dx \geq 0 \tag{5.1}$$

$$\exists x : SD2[x] > 0 \tag{5.2}$$

Figure 1, from [2], demonstrates an example in which the condition for a strategy stochastically dominating is not met, so the strategy considered is not consistently better results-wise than a randomly selected one. By contrast, Figure 2 shows an example of a case in which this domination does occur and the optimal strategy is considered to provide superior returns to a randomly chosen one.

6. Unique Features of CSCV

The primary driver for the choice of CSCV over K-FCV is inconsistencies between the sets created by K-FCV and those optimal for calculation of the Sharpe Ratio (and potentially other performance metrics). As demonstrated by the SR Confidence Bands defined in Definition 2.7, set size k must be small for $\widehat{SR}$ to be reliable. However, such a condition for K-FCV results in what [2] refers to as a "'hold-out' method, which [is] unreliable." LOOCV results in $k = T$, where T is the total size of the sample set. When this is the case, we do not have any reliable performance metric, as far as has yet been developed. When we consider alternative model types, we consider the possibility that it may be possible to develop

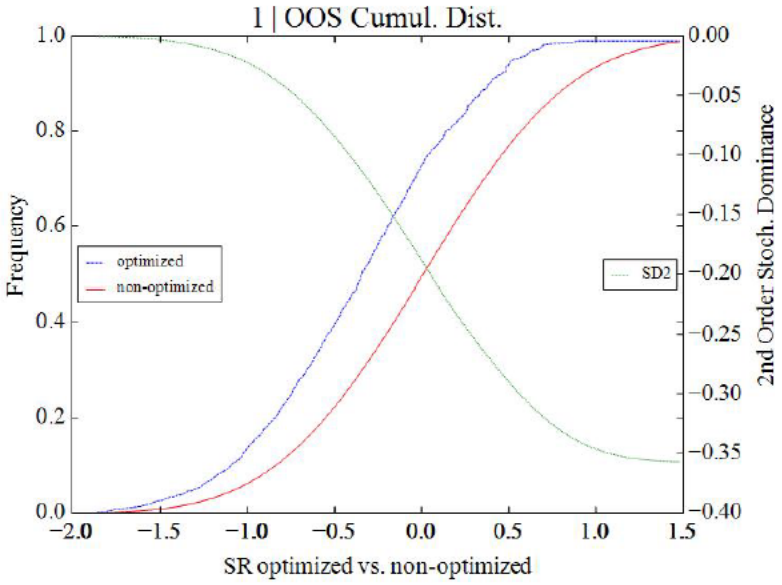

FIGURE 1. Stochastic dominance.

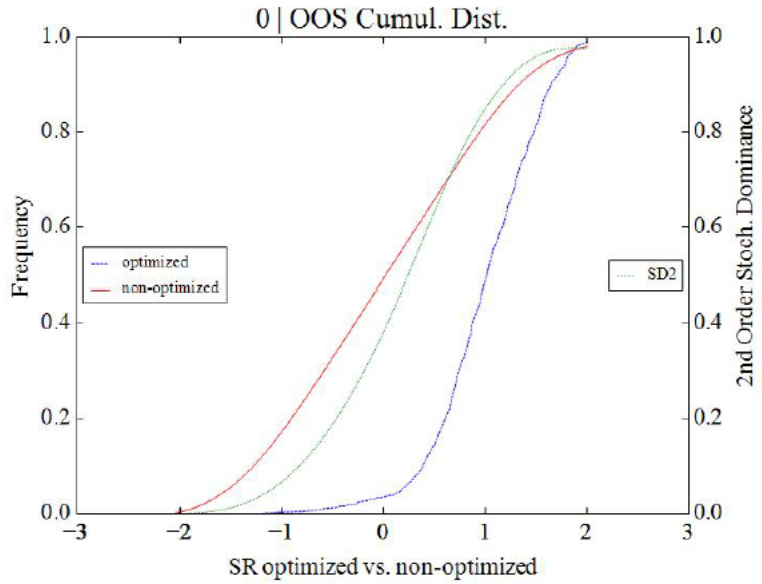

FIGURE 2. Stochastic dominance (example 2).

independent models which provide performance estimates based on single points, which could potentially allow this case. However, the advantages of avoiding this through a model-free system are later referenced by [3] as another advantage of CSCV.

A further advantage posed by CSCV is that it guarantees training and testing sets to be of the same size, guaranteeing the calculated accuracies of IS and OOS SRs to be comparable. Furthermore, in a similar realm, CSCV can be said to be symmetric, in that every testing or training set is also used as the other. This guarantees that performance decline over training/indicated by backtest, will only result from overfitting, the problem we're considering, rather than arbitrary set discrepancies.

Next, we note that CSCV does not utilize random allocations of subsample allocation, so respects time-dependence and season-dependent features. We also have replicability of results, as the logit distribution produced is deterministically dependent on the inputs, without randomness. Since the construction of the logit distribution is directly related to how robust a strategy selection procedure is, we can further state that the OOS performance rankings are consistent between trials. Finally, as mentioned above, since the PBO is model-free, the user typically does not need to forecast any parameters. The most important implication of this and the associated statement that the PBO is similarly non-parametric is that it provides a baseline of expectation of the potential case of backtesting not providing OOS performance insight. This thus allows us to specifically seek the case in which the logit distribution is primarily positive, with marginal coverage of the negative values, as a representation of positive association between backtesting and good OOS performance. However, in contrast to this, if model configuration is determined by a forecasting specification, this must be done with $T/2$ observations, where T is the samples provided to CSCV.

7. Accuracy of CSCV Algorithm

As CSCV analyzes an presents a probabilistic parameter, it is difficult to *a priori* state the correctness of the algorithm. As such, in an attempt to justify its proposal, [1] puts forth two methods of assessing whether PBO actually does serve as an indicator of underperformance of the median of OOS trials by the given security: Monte Carlo simulations and an application of the Extreme Value Theory. The former is a purely computational sampling technique, so we primarily consider the latter. We first recall the above proofs that $\widehat{SR}$ go to a Gaussian distribution asymptotically and that, per Proposition 1, a Gumbel distribution approximates the maximum IS performance of a strategy between N alternative backtests, given the Fisher-Tippet-Gnedenko Theorem. We follow the procedure from [1] for empirical confirmation and extension of these ideas.

We first consider a set of backtests with $N = 100, T = 1000, SR_n = 0$ given $n = 1, ..., N-1, SR_N = \widetilde{SR} > 0$. We first select 2 sets of equal size for IS and OOS, and select a strategy with $SR_n = 0$ when its IS SR exceeds that of the strategy with $SR_n = \widetilde{SR}$. Per global constraint of the SR by re-scaling and re-centering, and $|IS| = |OOS|$, $SR^*_{OOS} =\approx SR - SR^*_{IS}$. We use 4 propositions for the estimate of PBO:

(1) μ of all SRs OOS is null, $Me[SR_{OOS}] = 0$
(2) For a selected strategy, $SR_n = 0 \rightarrow SR^*_{OOS} \approx -max_N < Me[SR_{OOS}] iif SR^*_{IS} > 0$
(3) Selecting $SR_N = \widetilde{SR} \rightarrow SR^*_{OOS} \approx \widetilde{SR} - SR^*_{IS}, E[SR^*_{OOS}] > Me[SR_{OOS}] iif SR^*_{IS} \in (-\infty, w\widetilde{SR})$ and $E[SR^*_{OOS}] \leq Me[SR_{OOS}] iif SR^*_{IS} \in [2\widetilde{SR}, \infty)$

(4) $V[SR_{IS}] = V[SR] = \frac{1+\frac{1}{2}SR^2}{T}, V[SR_{OOS}] = 0$

We want that the maximum IS SR strategy performs below the median of OOS SR. We determine the specific Gumbel distribution with a variety of statements: $msa_N = max(\{SR_n | n = 1, ..., N-1\})$ with SR_n backtest-estimated during the nth trial. By the Maximum domain of Attraction of the Gumbel distribution, $max_N \sim \wedge[\alpha, \beta]$, α, β normalizing constants, $\wedge$ Gumbel distribution CDF. We have

$$E[max_N] = \alpha + \gamma\beta \tag{7.1}$$

$$\sigma[max_N] = \frac{\beta\pi}{\sqrt{6}} \tag{7.2}$$

With an estimate of $\hat{\sigma}[max_N]$, $\hat{\beta} = \frac{\hat{\sigma}[max_N]\sqrt{6}}{\pi}$
With this $\hat{\beta}$, we estimate $\hat{E}[max_N]$ with $\hat{a} = \hat{E}[max_N] - \gamma\hat{\beta}$
Finally, the probability we target $\phi = \phi_1 + \phi_2$:

$$\phi_1 = \int_{-\infty}^{2\widetilde{SR}} N\left[SR, \widetilde{SR}, \frac{1+\frac{1}{2}\widetilde{SR}^2}{T}\right](1 - \wedge[max(0, SR), \alpha, \beta])dSR \tag{7.3}$$

$$\phi_2 = \int_{2\widetilde{SR}}^{\infty} N\left[SR, \widetilde{SR}, \frac{1+\frac{1}{2}\widetilde{SR}^2}{T}\right]dSR \tag{7.4}$$

Note that the upper bound of $2\widetilde{SR}$ is chosen because, above that point, $SR^*_{OOS} < Me[SR_{OOS}]$. This is accounted for by phi_2, while phi_1 accounts for the other special case above. [1] cites a snippet of Python code which we do not reproduce here, but we will discuss their results and conclusions.

7.1. **Empirical Study of Accuracy of PBO Calculation.** Having established the EVT benchmark, probabilities for both it and the undescribed Monte Carlo method are laid out in the table below, in comparison to the mean and standard deviation produced by a series of iterations of CSCV. We note that the MC results were quite similar to those produced by EVT, so utilizing one or the other serves as an appropriate point of comparison. In analysis of the data, [1] cites that the average absolute error between Mean-CSCV and the EVT result is 2.1%, with standard deviation 2.9%. The maximum absolute error is 9.9%, for $\widetilde{SR} = 3, T = 500, N = 500$, with a more conservative estimate given by CSCV: 24.7% instead of 14.8%. The only case of underestimation of PBO by CSCV was with an absolute error of 0.1%. Total median error was 0.7%, which is small enough in consideration to state CSCV to be an accurate method of PBO calculation.

8. Conclusion

We first consider that there are a number of potential limitations and misuses of CSCV, including the fact that, depending on the complexity of certain strategies, symmetry of IS and OOS sets can be problematic, and the fact that the performance measure utilized is assumed to be equally weighted, and thus does not necessarily cater to measures dependent on forecasting equations and/or weighting. In application, there are 5 concerns outlined by the authors:

SR_Case	T	N	Mean_CSCV	Std_CSCV	Prob_MC	Prob_EVT	CSCV-EVT
0	500	500	1.000	0.000	1.000	1.000	0.000
0	1000	500	1.000	0.000	1.000	1.000	0.000
0	2500	500	1.000	0.000	1.000	1.000	0.000
0	500	100	1.000	0.000	1.000	1.000	0.000
0	1000	100	1.000	0.000	1.000	1.000	0.000
0	2500	100	1.000	0.000	1.000	1.000	0.000
0	500	50	1.000	0.000	1.000	1.000	0.000
0	1000	50	1.000	0.000	1.000	1.000	0.000
0	2500	50	1.000	0.000	1.000	1.000	0.000
0	500	10	1.000	0.001	1.000	1.000	0.000
0	1000	10	1.000	0.000	1.000	1.000	0.000
0	2500	10	1.000	0.000	1.000	1.000	0.000
1	500	500	0.993	0.007	0.991	0.994	-0.001
1	1000	500	0.893	0.032	0.872	0.870	0.023
1	2500	500	0.561	0.022	0.487	0.476	0.086
1	500	100	0.929	0.023	0.924	0.926	0.003
1	1000	100	0.755	0.034	0.743	0.713	0.042
1	2500	100	0.371	0.034	0.296	0.288	0.083
1	500	50	0.870	0.031	0.878	0.859	0.011
1	1000	50	0.666	0.035	0.628	0.626	0.041
1	2500	50	0.288	0.047	0.199	0.220	0.068
1	500	10	0.618	0.054	0.650	0.608	0.009
1	1000	10	0.399	0.054	0.354	0.360	0.039
1	2500	10	0.123	0.048	0.093	0.086	0.036
2	500	500	0.679	0.037	0.614	0.601	0.079
2	1000	500	0.301	0.038	0.213	0.204	0.097
2	2500	500	0.011	0.011	0.000	0.002	0.009
2	500	100	0.488	0.035	0.413	0.405	0.084
2	1000	100	0.163	0.045	0.098	0.099	0.065
2	2500	100	0.004	0.006	0.002	0.001	0.003
2	500	50	0.393	0.040	0.300	0.312	0.081
2	1000	50	0.113	0.044	0.068	0.066	0.047
2	2500	50	0.002	0.004	0.000	0.000	0.002
2	500	10	0.186	0.054	0.146	0.137	0.049
2	1000	10	0.041	0.027	0.011	0.023	0.018
2	2500	10	0.000	0.001	0.000	0.000	0.000
3	500	500	0.247	0.043	0.174	0.148	0.099
3	1000	500	0.020	0.017	0.005	0.005	0.015
3	2500	500	0.000	0.000	0.000	0.000	0.000
3	500	100	0.124	0.042	0.075	0.068	0.056
3	1000	100	0.007	0.008	0.001	0.002	0.005
3	2500	100	0.000	0.000	0.000	0.000	0.000
3	500	50	0.088	0.037	0.048	0.045	0.043
3	1000	50	0.004	0.006	0.002	0.001	0.003
3	2500	50	0.000	0.000	0.000	0.000	0.000
3	500	10	0.028	0.022	0.010	0.015	0.013
3	1000	10	0.001	0.002	0.000	0.001	0.000
3	2500	10	0.000	0.000	0.000	0.000	0.000

FIGURE 3. CSCV accuracy

(1) Occasionally, it is mathematically sensible for a researcher to remove certain trials from analysis in order to consider particular factors or such. Due to the structure of CSCV, this technique cannot here be utilized.
(2) No claims are made regarding backtest correctness, so this is a verification problem otherwise left to researchers.

(3) Since financial models often must consider possible structural breaks, those outside of a provided dataset are often a concern for researchers. Though those within a dataset are considered by CSCV, no account is taken of true OOS structural breaks.
(4) Though overfitting is posed as an important problem to battle in strategy choice, there do indeed exist high PBO strategies that remain skillfull and effective.
(5) As with any metric, utilizing CSCV as a guide to strategy development or search will result in misapplication of the probability. PBO is structured as an evaluation, not an effective objective function.

In conclusion, the works reviewed here have posed general frameworks for determination of PBO as a generic, symmetric, model-free, non-parametric assessment of underperformance of a highly fitted model IS when taken OOS. Though consideration of regression overfitting is commonplace in the field, these works are the first to consider the same concern for backtesting investment simulations. Doing so is necessary to consider the fact that backtesting provides memory in the process, which potentially hides problematic consequences. Furthermore, CSCV provides metrics for performance degredation, probability of loss, and stochastic dominance by a selected strategy. Ultimately, the hope of this proposal is to encourage consideration and control of both PBO and MinBTL (which implies that a greater number of trials should demand a higher IS Sharpe Ratio) in reporting of computation and backtest results among financial engineers, in order to add some degree of mathematical rigor to strategy choice.

References

1. Bailey, David H. and Borwein, Jonathan M. and Lopez de Prado, Marcos and Zhu, Qiji Jim, Mathematical Appendices to: 'The Probability of Backtest Overfitting' (February 22, 2015). Journal of Computational Finance (Risk Journals), 2015, Forthcoming. Available at SSRN: https://ssrn.com/abstract=2568435 or http://dx.doi.org/10.2139/ssrn.2568435
2. Bailey, David H. and Borwein, Jonathan M. 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 M. 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), Mai 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 Sharpe Ratio Efficient Frontier (April 2012). Journal of Risk, Vol. 15, No. 2, Winter 2012/13. Available at SSRN: https://ssrn.com/abstract=1821643 or http://dx.doi.org/10.2139/ssrn.1821643
5. 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

Department of Mathematics, University of Washington, Seattle, Washington 98105
Current address: Department of Mathematics, University of Washington, Seattle, Washington 98105
E-mail address: ac2zoom@cs.washington.edu