THE PROBABILITY OF BACKTEST OVERFITTING

David H. Bailey [α]

Jonathan M. Borwein [β]

Marcos López de Prado [γ]

Jim Zhu [δ]

First version: August 2013
This version: January 2014

[α] David H. Bailey is recently retired from the Lawrence Berkeley National Laboratory and is a Research Fellow at University of California, Davis, Department of Computer Science, Davis, CA, 95616 USA. E-mail: david@davidhbailey.com. URL: www.davidhbailey.com

[β] Jonathan M. Borwein is Laureate Professor of Mathematics at University of Newcastle (Australia), and a Fellow of the Royal Society of Canada, the Australian Academy of Science and of the AAAS. E-mail: jonathan.borwein@newcastle.edu.au. URL: www.carma.newcastle.edu.au/jon

[γ] Marcos López de Prado is Head of Quantitative Trading & Research at Hess Energy Trading Company, and a Research Affiliate at Lawrence Berkeley National Laboratory. E-mail: lopezdeprado@lbl.gov. URL: www.QuantResearch.info

[δ] Jim Zhu is Professor of Mathematics at Western Michigan University. E-mail: zhu@math-stat.wmich.edu. URL: http://homepages.wmich.edu/~zhu/

We are grateful to Tony Anagnostakis (Moore Capital), Marco Avellaneda (Courant Institute, NYU), Peter Carr (Morgan Stanley, NYU), David Donoho (Stanford University), Paul Embrechts (ETH Zürich), Matthew D. Foreman (University of California, Irvine), Ross Garon (SAC Capital), Jeffrey Lange (Guggenheim Partners), Attilio Meucci (KKR, NYU), Natalia Nolde (University of British Columbia and ETH Zürich), Riccardo Rebonato (PIMCO, University of Oxford), Joseph Romano (Stanford University) and Victoria Stodden (Stanford University).

Supported in part by the Director, Office of Computational and Technology Research, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy, under contract number DE-AC02-05CH11231.

THE PROBABILITY OF BACKTEST OVERFITTING

ABSTRACT

Most firms and portfolio managers rely on backtests (or historical simulations of performance) to select investment strategies and allocate them capital. Standard statistical techniques designed to prevent regression overfitting, such as hold-out, tend to be unreliable and inaccurate in the context of investment backtests. We propose a framework that estimates the *probability of backtest overfitting* (PBO) specifically in the context of investment simulations, through a numerical method that we call *combinatorially symmetric cross-validation* (CSCV). We show that CSCV produces accurate estimates of the probability that a particular backtest is overfit.

Keywords: Backtest, historical simulation, probability of backtest overfitting, investment strategy, optimization, Sharpe ratio, minimum backtest length, performance degradation.

JEL Classification: G0, G1, G2, G15, G24, E44.
AMS Classification: 91G10, 91G60, 91G70, 62C, 60E.

"This was our paradox: No course of action could be determined by a rule, because every course of action can be made to accord with the rule."

Ludwig Wittgenstein [1953]

1. INTRODUCTION

Despite of its limitations, the Sharpe ratio (SR) is the "gold standard" of investment performance evaluation. Bailey and López de Prado [2012] developed methodologies to assess the probability that a SR is inflated (PSR), and the minimum track record length (MinTRL) required for a SR to be statistically significant. These statistics were developed to judge the reliability of SR computed on live performance (track record). We have not been able to find similar statistics or methodologies applicable to judging backtested SR. Thus began our quest for a general methodology to assess the reliability of a backtest.

Perhaps the most common approach among practitioners is to require the researcher to "hold-out" a part of the available sample (also called "test set" method). This "hold-out" is used to estimate the OOS performance, which is then compared with the IS performance. If they are congruent, the investor has no grounds to "reject" the hypothesis that the backtest is overfit. The main advantage of this procedure is its simplicity. However, this approach is unsatisfactory for multiple reasons. First, if the data is publicly available, it is quite likely that the researcher has used the "hold-out" as part of the IS. Second, even if no "hold-out" data was used, any seasoned researcher knows well how financial variables performed over the OOS interval, and that information will be used in the strategy design, consciously or not (see Schorfheide and Wolpin [2012]). Third, hold-out is clearly inadequate for small samples. The IS will be too short to fit, and the OOS too short to conclude anything with sufficient confidence. Weiss and Kulikowski [1991] argue that hold-out should not be applied to an analysis with less than 1000 observations. For example, if a strategy trades on a weekly basis, hold-out could not be used on backtests of less than 20 years. In the same line, Van Belle and Kerr [2012] point out the high variance of hold-out's estimation errors. If we are unlucky, the chosen "hold-out" may be the one that refutes a valid strategy or supports an invalid strategy. Different "hold-outs" are likely to lead to different conclusions. Fourth, even if the researcher counts with a large sample, the OOS analysis will consume a large amount of the sample to be conclusive, which is detrimental to the strategy's design (see Hawkins [2004]). If the OOS is taken from the end of a time series, we are losing the most recent observations, which often are the most representative going forward. If the OOS is taken from the beginning of the time series, the testing will be done on the least representative portion of the data. Fifth, as long as the researcher tries more than one strategy configuration, overfitting is always present (see Bailey et al. [2013] for a proof). The hold-out method does not take into account the number of trials attempted before selecting a particular strategy configuration, and consequently hold-out cannot correctly assess a backtest's representativeness. Hold-out leaves the investor guessing to what degree the backtest is overfit. The answer to the question "is this backtest overfit?" is not a true or false, but a non-null probability that depends on the number of trials involved (an input ignored by hold-out). In this paper we will show a way to compute this probability.

A second approach popular among practitioners consists in modeling the underlying financial variable, generate random scenarios and measure the performance of the investment strategy on those scenarios. This presents the advantage of generating a distribution of outcomes, rather than relying on a single OOS performance estimate, as the "hold-out" method does. The disadvantages are that the model that generates random series of the underlying variable may also be overfit, may not contain all relevant statistical features, and it has to be customized to every variable (large development costs). Some retail trading platforms offer backtesting procedures based on this approach, such as random generation of tick data by fractal interpolation.

Several procedures have been proposed to determine whether an econometric model is overfit, see White [2000], Romano et al. [2005], Harvey et al. [2013] for a discussion in the context of Econometric models. Essentially these methods propose a way to adjust the p-values of estimated regression coefficients to account for the multiplicity of trials. These are valuable approaches when the trading rule relies on an econometric specification. That is not generally the case, as discussed in Bailey et al. [2013].

A generic, model-free and non-parametric approach to backtest overfitting would be useful. The main innovation of this paper is to propose a general framework that adapts recent advances in experimental mathematics, machine learning and decision theory to the very particular problem of assessing the representativeness of a backtest. This is not an easy problem, as evidenced by the lack of academic papers addressing a dilemma that most investors face. This gap in the literature is surprising, considering practitioners' reliance on backtests. One advantage of our solution is that it only requires time series of backtested performance. We avoid the credibility problem (of preserving a truly OOS test-set) by not requiring a fixed "hold-out," and swapping all IS and OOS sets. Our approach is generic in the sense of not requiring knowledge of the trading rule or forecasting equation. The output is a bootstrapped distribution of OOS Sharpe ratios, as well as a measure of the representativeness of the backtest that we call *probability of backtest overfitting* (PBO). Although in our examples we always choose the Sharpe ratio to evaluate performance, our methodology can be applied to any other performance measure.

The rest of the study is organized as follows: Section 2 sets the foundations of our framework. Section 3 defines the PBO and some other useful statistics that can be derived from this approach. Section 4 discusses some of the features of our framework, and how it relates to other machine learning methods. Section 5 lists some of the limitations of this method. Section 6 presents several test cases to illustrate how the PBO compares to different scenarios. Section 7 assesses the accuracy of our method using two alternative approaches: Monte Carlo and Extreme Value Theory. Section 8 discusses a practical application. Section 9 summarizes our conclusions. The mathematical appendices prove the propositions presented throughout the paper.

2. THE FRAMEWORK

2.1. DEFINITION OF OVERFITTING IN THE CONTEXT OF STRATEGY SELECTION

We ask the question: What is the probability that an “optimal” strategy is overfit? Intuitively, for overfitting to occur, the strategy configuration that delivers maximum performance IS must systematically underperform the rest of configurations OOS. The reason for this underperformance is that the IS “optimal” strategy is too closely tied to the training set, to the point that optimizing becomes detrimental.

More formally, we define the following probability space:

- Let be $R: \mathbb{R}^T \rightarrow \mathbb{R}$ a function that given the *n*th strategy’s performance series (of size *T*) returns a real-valued measure of performance R_n. For example, R could be the Sharpe ratio function applied on a vector of profit and losses (P&L) m_τ from a particular strategy.
- Given a set of N strategies, let be Ω the sample space of all possible strategy rankings according to R_n (in ascending order). Ω contains outcomes accounting for *N!* permutations of rankings.
- Let be $\mathcal{F}$ the set of events, which contains the sigma-algebra of Ω. In particular, one of its elements is the subset of all outcomes in Ω where the *n*th strategy ranked in the first half (i.e., below the median). Let us call this event $\underline{\mathcal{F}_n} \in \mathcal{F}, \forall n = 1, \dots, N$.
- Let be *Prob* a probability measure on $\mathcal{F}$. For example, the probability that the *n*th strategy’s performance is below the median of all performances corresponds to the relative frequency at which any outcome within event $\underline{\mathcal{F}_n}$ occurs.

Given the above probability space $(\Omega, \mathcal{F}, Prob)$, we can give a definition of backtest overfitting, in the context of investment strategy selection.

<u>DEFINITION 1 (Backtest Overfitting):</u> Let be n^* the strategy with optimal performance IS, i.e. $R_{n^*} \geq R_n, \forall n = 1, \dots, N$. Denote $\overline{R_{n^*}}$ the performance OOS of n^*. Let be $Me\left[\overline{R}\,\right]$ the median performance of all strategies OOS. Then, we say that a strategy selection process overfits if for a strategy n^* with the highest rank IS,

$$E\left[\overline{R_{n^*}}\,\right] < Me\left[\overline{R}\,\right] \tag{1}$$

<u>DEFINITION 2 (Probability of Backtest Overfitting):</u> Let be n^* the strategy with optimal performance IS. Because strategy n^* is not necessarily optimal OOS, there is a non-null probability that $\overline{R_{n^*}} < Me\left[\overline{R}\,\right]$. We define the probability that the selected strategy n^* is overfit as

$$PBO \equiv Prob\left[\overline{R_{n^*}} \; < Me\left[\overline{R}\,\right]\right] \tag{2}$$

In other words, we say that a strategy selection process overfits if the expected performance of the strategies selected IS is less than the median performance OOS of all strategies. In that situation, the strategy selection process becomes in fact detrimental. Note that in this context IS corresponds to the subset of observations m_τ used to select the optimal strategy n^* among the *N* alternatives. With IS we do not mean the period on which the investment model underlying the strategy was estimated (e.g., the period on which crossing moving averages are computed, or a forecasting regression model is estimated). Consequently, in the above definition we refer to overfitting in relation to the strategy selection process, not a strategy's model calibration (e.g., in the context of regressions). That is the reason we were able to define overfitting without knowledge of the strategy's underlying models, i.e. in a model-free and non-parametric manner. Next, we describe a procedure to determine whether overfitting is taking place in strategy selection, based on this logic.

2.2. THE PROCEDURE

Suppose that a researcher is developing an investment strategy. She considers a family of system specifications and parametric values to be backtested, in an attempt to uncover the most profitable incarnation of that idea. For example, in a trend following moving average strategy, the researcher could try alternative sample lengths on which the moving averages are computed, entry thresholds, exit thresholds, stop losses, holding periods, sampling frequencies, etc. As a result, the researcher ends up running a number *N* of alternative model configurations (or trials), out of which one is chosen according to some performance evaluation criterion, such as the Sharpe ratio.

First, we form a matrix ***M*** by collecting the performance series from the *N* trials. In particular, each column *n=1,…,N* represents a vector of P&L (profits and losses) over *t=1,…,T* observations associated with a particular model configuration tried by the researcher. ***M*** is therefore a real-valued matrix of order (TxN). The only conditions we impose are that: i) ***M*** is a true matrix, i.e. with the same number of rows for each column, where observations are synchronous for every row across the *N* trials, and ii) the performance evaluation metric used to choose the "optimal" strategy can be estimated on subsamples of each column. For example, if that metric was the Sharpe ratio, we would expect that the IID Normal distribution assumption can be held on various slices of the reported performance. If different model configurations trade with different frequencies, observations are aggregated to match a common index *t=1,…,T*.

Second, we partition ***M*** across rows, into an even number *S* of disjoint submatrices of equal dimensions. Each of these submatrices $\boldsymbol{M_s}$, with *s=1,…,S*, is of order $\left(\frac{T}{S}xN\right)$.

Third, we form all combinations C_S of $\boldsymbol{M_s}$, taken in groups of size $\frac{S}{2}$. This gives a total number of combinations

$$\binom{S}{S/_2} = \binom{S-1}{S/_2 - 1}\frac{S}{S/_2} = \cdots = \prod_{i=0}^{S/_2 - 1} \frac{S-i}{S/_2 - i} \tag{3}$$

For instance, if $S{=}16$, we will be forming 12,780 combinations. Each combination $c \in C_S$ is composed of ${}^{S}\!/_{2}$ submatrices $\boldsymbol{M_s}$.

Fourth, for each combination $c \in C_S$, we:

a) Form the *training set* $\boldsymbol{J}$, by joining the ${}^{S}\!/_{2}$ submatrices $\boldsymbol{M_s}$ that constitute *c*. $\boldsymbol{J}$ is a matrix of order $\left(\frac{T}{S}\frac{S}{2}xN\right) = \left(\frac{T}{2}xN\right)$.

b) Form the *testing set* $\bar{\boldsymbol{J}}$, as the complement of $\boldsymbol{J}$ in $\boldsymbol{M}$. In other words, $\bar{\boldsymbol{J}}$ is the $\left(\frac{T}{2}xN\right)$ matrix formed by all rows of $\boldsymbol{M}$ that are not part of $\boldsymbol{J}$.

c) Form a vector $\boldsymbol{R}$ of performance statistics of order *N*, where the *n*-th item of $\boldsymbol{R}$ reports the performance associated with the *n*-th column of $\boldsymbol{J}$ (the training set).

d) Determine the element n^* such that $R_n \leq R_{n^*}$, $\forall n = 1, \dots, N$. In other words, $n^* = \arg max_n \{R_n\}$.

e) Form a vector $\bar{\boldsymbol{R}}$ of performance statistics of order *N*, where the *n*-th item of $\bar{\boldsymbol{R}}$ reports the performance associated with the *n*-th column of $\bar{\boldsymbol{J}}$ (the testing set).

f) Determine the relative rank of $\overline{R_{n^*}}$ within $\bar{\boldsymbol{R}}$. We will denote this relative rank as $\bar{\omega}_c$, where $\bar{\omega}_c \in (0,1)$. This is the relative rank of the OOS performance associated with the trial chosen IS. If the strategy optimization procedure is not overfitting, we should observe that $\overline{R_{n^*}}$ systematically outperforms $\bar{\boldsymbol{R}}$ (OOS), just as R_{n^*} outperformed $\boldsymbol{R}$.

g) We define the logit $\lambda_c = Ln\frac{\bar{\omega}_c}{1-\bar{\omega}_c}$. This presents the property that $\lambda_c = 0$ when $\overline{R_{n^*}}$ coincides with the median of $\bar{\boldsymbol{R}}$. High logit values imply a consistency between IS and OOS performance, which indicates a low level of backtest overfitting.

Fifth, compute the distribution of ranks OOS by collecting all the λ_c, for $c \in C_S$. $f(\lambda)$ is then the relative frequency at which λ occurred across all C_S, with $\int_{-\infty}^{\infty} f(\lambda) d\lambda = 1$. Table 1 schematically represents how the combinations in C_S are used to produce training and testing sets, where $S{=}4$.

[TABLE 1 HERE]

3. OVERFIT STATISTICS

The framework introduced in Section 2 allows us to characterize the reliability of a strategy's backtest in terms of four complementary analyses:

1. Probability of Backtest Overfitting (PBO): The probability that the model configuration selected as optimal IS will underperform the median of the *N* model configurations OOS.
2. Performance degradation: It determines to what extent greater performance IS leads to lower performance OOS, an occurrence associated with the memory effects discussed in Bailey et al. [2013].

3. Probability of loss: The probability that the model selected as optimal IS will deliver a loss OOS.
4. Stochastic dominance: This analysis determines whether the procedure used to select a strategy IS is preferable to randomly choosing one model configuration among the *N* alternatives.

3.1. PROBABILITY OF OVERFITTING (PBO)

PBO was defined earlier as $Prob\left[\overline{R_{n^*}} < Me\left[\overline{R}\right]\right]$, and can be estimated as $\phi = \int_{-\infty}^{0} f[\lambda]d\lambda$. This represents the rate at which optimal IS strategies underperform the median of the OOS trials. The analogue of $\overline{\boldsymbol{R}}$ in medical research is the placebo given to a portion of patients in the test set. If the backtest is truly helpful, the optimal strategy selected IS should outperform most of the *N* trials OOS. That is the case when $\lambda_c > 0$. There are three relevant scenarios associated with ϕ:

- $\phi \approx 0$: A low proportion of combinations *C* exhibited $\lambda_c < 0$. Thus, the optimal IS strategy outperformed the median of trials in most of the testing sets. There is no significant overfitting, because choosing the optimal strategy IS indeed helped improve performance OOS.
- $\phi = \frac{1}{2}$: In half of the combinations *C* it occurred that $\lambda_c < 0$. The optimal IS strategy underperformed in as many trials as it outperformed. The expected performance of the optimal strategy identified by the backtest equals the median performance from the trials. Backtests are overfit to the point that the strategy selection procedure does not add value.
- $\phi \gg \frac{1}{2}$: In more than half of the combinations *C* it occurred that $\lambda_c < 0$. Thus, the optimal IS strategy underperformed the median of trials in more than half of the testing sets. The degree of overfitting is so prevalent that choosing the optimal strategy leads to worse expected performance than picking a strategy at random from the trials.

3.2. PERFORMANCE DEGRADATION AND PROBABILITY OF LOSS

Section 2.2 introduced the procedure to compute, among other results, the pair $\left(R_{n^*}, \overline{R_{n^*}}\right)$ for each combination $c \in C_S$. Note that while we know that R_{n^*} is the maximum of $\boldsymbol{R}$, $\overline{R_{n^*}}$ is not necessarily the maximum of $\overline{\boldsymbol{R}}$. Because we are trying every combination of $\boldsymbol{M_s}$ taken in groups of size $\frac{S}{2}$, there is no reason to expect the distribution of $\boldsymbol{R}$ to dominate over $\overline{\boldsymbol{R}}$. The implication is that, generally, $\overline{R_{n^*}} < max\{\overline{\boldsymbol{R}}\} \approx max\{\boldsymbol{R}\} = R_{n^*}$. For a regression $\left[\overline{R_{n^*}}\right]_c = \alpha + \beta[R_{n^*}]_c + \varepsilon_c$, the β will be negative in most practical cases, due to compensation effects described in Bailey et al. [2013]. An intuitive explanation for this negative slope is that overfit backtests minimize future performance: The model is so fit to the past, that it is rendered unfit for the future. And the more overfit a backtest is, the more memory is accumulated against its future performance.

It is interesting to plot the pairs $\left(R_{n^*}, \overline{R_{n^*}}\right)$ to visualize how strong is the performance degradation, and obtain a more realistic range of attainable performance OOS (see Figure

8). A particularly useful statistic is the proportion of combinations with negative performance, $Prob\left[\left[\overline{R_{n^*}}\right]_c < 0\right]$. Note that, even if $\phi \approx 0$, $Prob\left[\left[\overline{R_{n^*}}\right]_c < 0\right]$ could be high, in which case the strategy's performance OOS is poor for reasons other than overfitting.

[FIGURE 1 HERE]

Figure 1-1 plots the performance degradation associated with the backtest of an investment strategy. We can once again appreciate the negative relationship between greater SR IS and SR OOS, which indicates that at some point seeking the optimal performance becomes detrimental. Whereas 100% of the SR IS are positive, about 78% of the SR OOS are negative. Also, SR IS range between 1 and 3, indicating that backtests with high Sharpe ratios tell us nothing regarding the representativeness of that result. We cannot hope escaping the risk of overfitting by exceeding some SR IS threshold. On the contrary, it appears that the higher the SR IS, the lower the SR OOS. In this example we are evaluating performance using the Sharpe ratio, however we stress that our procedure is generic and can be applied to any performance evaluation metric R (Sortino ratio, Jensen's Alpha, Probabilistic Sharpe Ratio, etc.).[1] Figure 1-2 shows the distribution of logits for the same strategy, with a PBO of 74%.

[FIGURE 2 HERE]

Figure 2-1 plots the performance degradation associated with the backtest of a real investment strategy. The regression line that goes through the pairs of (SR IS, SR OOS) is much less steep, and only 3% of the SR OOS are negative. Figure 2-2 shows the distribution of logits, with a PBO of 0.04%. According to this analysis, it is unlikely that this backtest is overfit. The chances that this strategy performs well OOS are much greater than in the previous example.

3.3. STOCHASTIC DOMINANCE

A further application of the results derived in Section 2.2 is to determine whether the distribution of $\overline{R_{n^*}}$ across all $c \in C_S$ stochastically dominates over the distribution of all $\overline{\boldsymbol{R}}$. Should that not be the case, it would present strong evidence that strategy selection optimization does not provide consistently better OOS results than a random strategy selection. One reason that makes the concept of stochastic dominance useful is that it allows us to rank gambles or lotteries without having to make strong assumptions regarding an individual's utility function. See Hadar and Russell [1969] for an introduction.

Let $\overline{X_{n^*}}$ represent a random draw from $\overline{R_{n^*}}$ across all $c \in C_S$, and X a random draw from vector $\overline{\boldsymbol{R}}$. In the context of our framework, first-order stochastic dominance occurs if $Prob\left[\overline{X_{n^*}} \geq x\right] \geq Prob\left[\overline{X} \geq x\right]$ for all x, and for some x, $Prob\left[\overline{X_{n^*}} \geq x\right] >$

[1] Although beyond the scope of this study, it would be interesting to carry out a detailed analysis of what performance evaluation metrics are less prone to overfitting, according to the statistics introduced in Sections 3.1, 3.2 and 3.3.

$Prob\left[\overline{X} \geq x\right]$. It can be verified visually by checking whether the cumulative distribution function of $\overline{X_{n^*}}$ is not above the cumulative distribution function of $\overline{\boldsymbol{X}}$ for all possible outcomes, and at least for one outcome the former is strictly below the latter. Under such circumstances, the decision maker would prefer the criterion used to produce $\overline{R_{n^*}}$ over a random sampling of $\overline{\boldsymbol{R}}$, as long as her utility function is weakly increasing.

A less demanding criterion is second-order stochastic dominance. This requires that $SD2[x] = \int_{-\infty}^{x}\left(Prob\left[\overline{X} \leq x\right] - Prob\left[\overline{X_{n^*}} \leq x\right]\right)dx \geq 0$ for all *x*, and that $SD2[x] > 0$ at some *x*. When that is the case, the decision maker would prefer the criterion used to produce $\overline{R_{n^*}}$ over a random sampling of $\overline{\boldsymbol{R}}$, as long as she is risk averse and her utility function is weakly increasing.

[FIGURE 3 HERE]

Figure 3 provides an example of the cumulative distribution function of $\overline{X_{n^*}}$ across all $c \in C_S$ (red line) and $\overline{X}$ (blue line), as well as the second order stochastic dominance ($SD2[x]$, green line) for every OOS SR. It has been computed on the same backtest used for Figure 1. Consistent with that result, the overall distribution of OOS performance dominates the OOS performance of the optimal strategy selection procedure, a clear sign of overfitting.

[FIGURE 4 HERE]

Figure 4 provides a counter-example, based on the same real investment strategy used in Figure 2. It indicates that the strategy selection procedure used in this backtest actually added value, as the distribution of OOS performance for the selected strategies clearly dominates the overall distribution of OOS performance. First-order stochastic dominance is a sufficient condition for second-order stochastic dominance, and the plot of $SD2[x]$ is consistent with that fact.

4. FEATURES OF THE PROPOSED FRAMEWORK

Our testing method combines multiple discoveries in the fields of machine learning (combinatorial optimization, jackknife, cross-validation) and decision theory (logistic function, stochastic dominance). Standard cross-validation methods include *k-fold cross-validation* (K-FCV) and *leave-one-out cross-validation* (LOOCV).

K-FCV randomly divides the sample of size *T* into *k* subsamples of size $\frac{T}{k}$. Then it sequentially tests on each of the *k* samples the model trained on the $T - \frac{T}{k}$ sample. Although a very valid approach in many situations, we believe that our procedure is more adequate than K-FCV in the context of strategy selection. In particular, we would like to compute the Sharpe ratio (or any other performance measure) on each of the *k* testing sets of size $\frac{T}{k}$. This means that *k* must be sufficiently small, so that the Sharpe ratio estimate is reliable (see Bailey and López de Prado [2012] for a discussion of Sharpe ratio confidence bands). But if k is small, K-FCV will essentially reduce to a "hold-out" method, which we have argued is inaccurate.

LOOCV is a K-FCV where $k = T$. We are not aware of any reliable performance metric computed on a single OOS observation.

The cross-validation method we have proposed in Section 2.2 differs from K-FCV and LOOCV. The key idea is to generate $\binom{S}{S/2}$ testing sets of size $\frac{T}{2}$ by recombining the S slices of the overall sample of size T. To facilitate the discussion, we will name it *combinatorially symmetric cross-validation* (CSCV). This procedure presents a number of advantages. First, CSCV ensures that the training and testing sets are of equal size, thus providing comparable accuracy to the IS and OOS Sharpe ratios (or any other performance metric that is susceptible to sample size). This is important, because making the testing set smaller than the training set (as hold-out does) would mean that we are evaluating with less accuracy OOS than the one used to choose the optimal strategy. Second, CSCV is *symmetric*, in the sense that all training sets are re-used as testing sets and vice versa. In this way, the decline in performance can only result from overfitting, not arbitrary discrepancies between the training and testing sets. Third, CSCV respects the time-dependence and other seasonalities present in the data, because it does not require a random allocation of the observations to the S subsamples. We avoid that requirement by recombining the S subsamples into the $\binom{S}{S/2}$ testing sets. Fourth, CSCV derives a non-random distribution of logits, in the sense that each logit is deterministically derived from one item in the set of combinations C_S. Similarly to jackknife resampling, running CSCV twice on the same inputs generates identical results. Therefore, for each analysis, CSCV will provide a single result, ϕ, which can be independently replicated and verified by another user. Fifth, the dispersion of the distribution of logits conveys relevant information regarding the robustness of the strategy selection procedure. A robust strategy selection leads to a consistent OOS performance rankings, which translate into similar logits.

Sixth, our procedure to estimate PBO is model-free, in the sense that it does not require the researcher to specify a forecasting model or the definitions of forecasting errors. It is also non-parametric, as we are not making distributional assumptions on PBO. This is accomplished by using the concept of logit, λ_c. A logit is the logarithm of odds. In our problem, the odds are represented by relative ranks (i.e., the odds that the optimal strategy chosen IS happens to underperform OOS). The logit function presents the advantage of being the inverse of the sigmoidal logistic distribution, which resembles the cumulative Normal distribution. As a consequence, if $\overline{\omega}_c$ are distributed close to uniform (the case when the backtest appears to be informationless), the distribution of the logits will approximate the standard Normal. This is important, because it gives us a baseline of what to expect in the threshold case where the backtest does not seem to provide any insight into the OOS performance. If good backtesting results are conducive to good OOS performance, the distribution of logits will be centered in a significantly positive value, and its left tail will marginally cover the region of negative logit values, making $\phi \approx 0$.

A key parameter of our procedure is the value of *S*. This regulates the number of submatrices $\boldsymbol{M_s}$ that will be generated, each of order $\left(\frac{T}{S} xN\right)$, and also the number of logit values that will be computed, $\begin{pmatrix} S \\ S/_2 \end{pmatrix}$. S must be large enough so that the number of combinations suffices to draw inference. If *S* is too small, the left tail of the distribution of logits will be underrepresented. On the other hand, if we believe that the performance series is time-dependent and incorporates seasonal effects, *S* cannot be too large, or the relevant time structure may be shuttered across the partitions. For example, if the backtest includes more than six years of data, *S=24* generates partitions spanning over a quarter each, which would preserve daily, weekly and monthly effects, while producing a distribution of 2,704,156 logits. In contrast, if we are interested in quarterly effects, we have two choices: i) Work with *S=12* partitions, which will give us 924 logits, and/or ii) double *T*, so that *S* does not need to be reduced.

Another key parameter is the number of trials (i.e., the number of columns in $\boldsymbol{M_s}$). Hold-out's disregard for the number of trials attempted was the reason we concluded it was an inappropriate method to assess a backtest's representativeness (see Bailey et al. [2013] for a proof). *N* must be large enough to provide sufficient granularity to the values of the relative rank, $\overline{\omega}_c$. If *N* is too small, $\overline{\omega}_c$ will only adopt a very discrete number of values, which will translate in a very discrete number of logits, making $f(\lambda)$ too discontinuous, adding estimation error to the evaluation of ϕ. For example, if the investor is sensitive to values of $\phi < \frac{1}{10}$, it is clear that the range of values that the logits can adopt must be greater than 10, and so $N \gg 10$ is required. Other considerations regarding *N* will be discussed in the following Section.

Finally, PBO is evaluated by comparing combinations of $\frac{T}{2}$ observations with their complementary. But the backtest counts with T observations, rather than only $\frac{T}{2}$. Therefore, T should be chosen to be double of the number of observations used by the investor to choose a model configuration or determine a forecasting specification.

5. LIMITATIONS AND MISUSE

This procedure was designed to evaluate PBO under minimal assumptions and input requirements. In doing so, we have attempted to provide a very general (in fact, model-free and non-parametric) procedure against which IS backtests can be benchmarked. The reader should understand that model-specific methods may be more accurate in certain instances, by sacrificing generality and comparability across models. For example, if a forecasting equation was used to generate the trials, it would be possible to develop a framework that evaluates PBO particular to that forecasting equation. Because many investment strategies lack a forecasting equation, this is not always an option. We believe that a general procedure is useful in the context of deciding across multiple investment options, as exemplified by the success of benchmark methods such as the Sharpe ratio. But like in the case of the Sharpe ratio, it is important to discuss the limitations of our approach.

First, the researcher must provide as many P&L series (*N*) as truly tested, and test as many strategy configurations as reasonable and feasible. Hiding trials will lead to an underestimation of the overfit, because each logit will be evaluated under a biased relative rank $\overline{\omega}_c$. It would be the equivalent to removing subjects from the trials of a new drug, once we have verified that the drug was not effective on them. Likewise, adding trials that are doomed to fail in order to make one particular model configuration succeed biases the result. If a model configuration is obviously flawed, it should have never been tried in the first place. This procedure aims at evaluating how reliable a backtest selection process is when choosing among feasible strategy configurations. As a rule of thumb, the researcher should backtest as many theoretically reasonable strategy configurations as possible.

Second, this procedure does nothing to evaluate the correctness of a backtest. If the backtest is flawed due to incorrect assumptions, such as transaction costs or using data not available at the moment of making a decision, our approach will be making an assessment based on flawed information.

Third, this procedure only takes into account structural breaks as long as they are present in the dataset of length *T*. If the structural break occurs outside the boundaries of the available dataset, the strategy may be overfit to a particular data regime, which our PBO has failed to account for because the entire set belongs to the same regime. This invites the more general warning that the dataset used for any backtest is expected to be representative of future states of the modeled financial variable.

6. TEST CASES

We will compare how PBO responds to several test cases: Full, high and low overfit. These cases are created by setting a matrix ***M*** with *N-1* trials of length *T* and null Sharpe ratio. If we add to that matrix ***M*** a trial with Sharpe ratio zero, the strategy selection procedure will deliver a full overfit, because selecting the strategy with highest Sharpe ratio IS in this context cannot improve performance OOS over the median (zero). If we add to ***M*** a trial with Sharpe ratio 1, selecting the strategy with the highest Sharpe ratio IS may still improve performance OOS over the median (zero), giving us the high overfit case. If we add to ***M*** a trial with Sharpe ratio 2, selecting the strategy with the highest Sharpe ratio IS will likely improve performance OOS over the median (zero), giving us the low overfit case.

6.1. TEST CASE 1: FULL OVERFIT

We create a matrix ***M*** where *T=1000*, *N=100* as follows.

1. For each trial, *n=1,...,100*:
    a. Form a vector of *T* random draws from a standard Normal distribution.
    b. Re-scale and re-centre the vector so that its Sharpe ratio is 0.

2. Combine the *N* vectors into a matrix ***M***.

If we choose IS the trial with highest Sharpe ratio, this cannot be expected to outperform OOS. If our procedure is accurate, we should obtain that $\phi \to 1$, indicating that in virtually all cases the "optimal" strategy IS happened to underperform OOS the median of trials. In effect, our simulations show that CSCV correctly determined that backtests are almost certain to overfit in this situation.

Increasing the sample size to *T=2000* has no effect, and we still obtain $\phi \to 1$. Similarly, increasing the number of trials to *N=200* still produces $\phi \to 1$.

6.2. TEST CASE 2: HIGH OVERFIT

We create a matrix ***M*** where *T=1000, N=100* as follows.

1. For each trial, *n=1,...,100*:
    a. Form a vector of *T* random draws from a standard Normal distribution.
    b. Re-scale and re-centre that vector so that its Sharpe ratio is 0.
2. Re-scale and re-centre the *Nth* vector so that its Sharpe ratio is 1.
3. Combine the *N* vectors into a matrix ***M***.

Because one of the trials is expected to succeed OOS, PBO is not 1. At the same time, a random sample with a Sharpe ratio of 0 over *T* observations is likely to produce IS a Sharpe ratio above 1 over $\frac{T}{2}$ observations (see Bailey et al. [2013]). Accordingly, it is very likely that the strategy selection procedure will choose one trial with a high Sharpe ratio IS, only to underperform OOS the median of trials. The conclusion is that PBO will still be high in this scenario. Our Monte Carlo simulations confirm that intuition, by computing overfitting probabilities between 0.7 and 0.8.

Increasing the number of trials to *N=200* slightly increases ϕ, which now ranges between 0.75 and 0.85. The reason is, as more trials are added, the risk of overfitting increases, thus the importance that the researcher reports all the possibilities truly tested.

Increasing the sample size to *T=2000* reduces PBO to values between 0.4 and 0.5, because the larger the number of available observations, the more informative is the performance IS. These empirical findings are consistent with Proposition 1 and Theorem 1, discussed in Bailey et al. [2013].

6.3. TEST CASE 3: LOW OVERFIT

We create a matrix ***M*** where *T=1000*, *N=100* as follows.

1. For each trial, *n=1,...,100*:
    a. Form a vector of *T* random draws from a standard Normal distribution.
    b. Re-scale and re-centre that vector so that its Sharpe ratio is 0.
2. Re-scale and re-centre the *Nth* vector so that its Sharpe ratio is 2.
3. Combine the *N* vectors into a matrix ***M***.

Given that one of the trials performs significantly better than the rest over T observations, we expect a greater probability of it being selected IS over $\frac{T}{2}$ observations. Still, there is a non-null probability that the strategy selection procedure fails to do so, because it is still possible for a subsample of that trial to underperform IS the subsample of one of the other trials, hence leading to overfitting. Accordingly, our simulations estimate overfitting probabilities ranging between 0.1 and 0.2.

Increasing the number of trials to $N=200$ slightly increases ϕ, which now ranges between 0.2 and 0.3. Increasing the sample size to $T=2000$ reduces PBO to values between 0 and 0.04, just as we argued in Bailey et al. [2013].

The reader may draw a parallel between these results and her knowledge of overfitting in regression models. As it is documented in most statistics textbooks, the probability of overfitting a regression model increases as the number of degrees of freedom decreases. Our $\boldsymbol{M}$ matrix has its regression analogue in the $\boldsymbol{X}$ matrix of explanatory (or exogenous) variables, shaped in T rows (time observations) and N factors (columns), used to fit some other explained variable. In such regression model, the number of degrees of freedom is T-N. The larger T is, the more degrees of freedom, and the lower the risk of overfitting. Conversely, the larger N is, the less degrees of freedom and the higher the risk of overfitting. In conclusion, ϕ behaves as we would have expected in the familiar case of regression models.

7. ACCURACY OF THE TEST

In the previous Section, as we increased the Sharpe ratio of the *Nth* trial above the other trials in matrix $\boldsymbol{M}$, we observed a decrease in PBO. Decreasing N and increasing T had a similar effect. Thus, overfitting estimates in the test cases above seem reasonable in an ordinal sense. The question remains, does the actual estimate correspond to the probability of the selected strategy to underperform the median of trials OOS? We evaluate the accuracy of our CSCV procedure to determine PBO in two different ways: Via Monte Carlo (MC) simulations, and applying Extreme Value Theory (EVT).

7.1. ACCURACY BY MC

In order to determine the MC accuracy of our PBO estimate, we have generated 1,000 matrices $\boldsymbol{M}$ (experiments) for various test cases of order $(TxN) = (1000x100)$, and computed the proportion of experiments that yielded an OOS performance below the median. If our CSCV procedure is accurate, the PBO that we estimated by resampling slices of a single matrix $\boldsymbol{M}$ should be close to the probability derived from generating 1,000 matrices $\boldsymbol{M}$ and computing the proportion of IS optima that underperformed the median OOS.

Snippet 1 lists a code written in Python that estimates PBO via Monte Carlo.

```
#!/usr/bin/env python
# On 20130704 by lopezdeprado@lbl.gov
#---------------------------------------------------------------------------------------
```

```
def testAccuracy_MC(sr_base,sr_case):
    # Test the accuracy of CSCV against hold-out
    # It generates numTrials random samples and directly computes the …
    # … proportion where OOS performance was below the median.
    length,numTrials,numMC=1000,100,1000
    pathOutput=‘H:/Studies/Quant #23/paper/’
    #1) Determine mu,sigma
    mu_base,sigma_base=sr_base/(365.25*5/7.),1/(365.25*5/7.)**.5
    mu_case,sigma_case=sr_case/(365.25*5/7.),1/(365.25*5/7.)**.5
    hist,probOverfit=[],0
    #2) Generate trials
    for m in range(numMC):
        for i in range(1,numTrials):
            j=np.array([gauss(0,1) for j in range(length)])
            j*=sigma_base/np.std(j) # re-scale
            j+=mu_base-np.mean(j) # re-center
            j=np.reshape(j,(j.shape[0],1))
            if i==1:pnl=np.copy(j)
            else:pnl=np.append(pnl,j,axis=1)
        #3) Add test case
        j=np.array([gauss(0,1) for j in range(length)])
        j*=sigma_case/np.std(j) # re-scale
        j+=mu_case-np.mean(j) # re-center
        j=np.reshape(j,(j.shape[0],1))
        pnl=np.append(pnl,j,axis=1)
        #4) Run test
        # Reference distribution
        mu_is=[np.average(pnl[:length/2,i]) for i in range(pnl.shape[1])]
        sigma_is=[np.std(pnl[:length/2,i]) for i in range(pnl.shape[1])]
        mu_oos=[np.average(pnl[length/2:,i]) for i in range(pnl.shape[1])]
        sigma_oos=[np.std(pnl[length/2:,i]) for i in range(pnl.shape[1])]
        sr_is=[mu_is[i]/sigma_is[i] for i in range(len(mu_is))]
        sr_oos=[mu_oos[i]/sigma_oos[i] for i in range(len(mu_oos))]
        print m,sr_is.index(max(sr_is)),max(sr_is), \
            sr_oos[sr_is.index(max(sr_is))]
        sr_oos_=sr_oos[sr_is.index(max(sr_is))]
        hist.append(sr_oos_)
        if sr_oos_<np.median(sr_oos):probOverfit+=1
    probOverfit/=float(numMC)
    print probOverfit
    return
```

Snippet 1 – Python code for estimating PBO via Monte Carlo

7.2. ACCURACY BY EVT

Backtesting a number *N* of alternative strategy configurations and selecting the trial that exhibits maximum performance IS sets the background for applying Extreme Value Theory. From Eq. (1) we know that Sharpe ratio estimates asymptotically follow a Gaussian distribution. Proposition 1 discussed the distribution of the maximum of *N* independent random variables. As part of its proof, we applied the Fisher-Tippet-Gnedenko theorem to the Gaussian distribution, and concluded that the maximum performance IS among *N* alternative backtests can be approximated through a Gumbel distribution.

More formally, suppose a set of backtests where *N=100*, *T=1000*, $SR_n = 0$ for *n=1,…,N-1* and $SR_N = \widetilde{SR} > 0$. The sample length is divided in two sets of equal size ($\frac{T}{2}$), IS and OOS. A strategy with $SR_n = 0$ is selected when its IS SR is greater than the IS SR of the strategy with $SR_N = \widetilde{SR}$. Because the Sharpe ratio has been globally constrained for the whole sample by re-scaling and re-centering, and IS has the same length as OOS, $SR^*_{OOS} \approx SR - SR^*_{IS}$. By virtue of this global constraint, the following propositions can be used to estimate PBO:

i. The median of all Sharpe ratios OOS is null, $Me\left[SR_{OOS}\right] = 0$.

ii. Selecting a strategy with $SR_n = 0$ leads to $SR^*_{OOS} \approx -max_N < Me\left[SR_{OOS}\right]$ *iif* $SR^*_{IS} > 0$.

iii. Selecting a strategy with $SR_N = \widetilde{SR}$ leads to $SR^*_{OOS} \approx \widetilde{SR} - SR^*_{IS}$, where $E[SR^*_{OOS}] > Me\left[SR_{OOS}\right]$ *iif* $SR^*_{IS} \in \left(-\infty, 2\widetilde{SR}\right)$ and $E[SR^*_{OOS}] \leq Me\left[SR_{OOS}\right]$ *iif* $SR^*_{IS} \in \left[2\widetilde{SR}, \infty\right)$.

iv. Computing SR_{IS} fully determines SR_{OOS} as a result of the global constraint. Thus, $V[SR_{IS}] = V[SR] = \frac{1+\frac{1}{2}SR^2}{T}$, and $V[SR_{OOS}] = 0$.

Our goal is to compute the probability that the strategy with maximum Sharpe ratio IS performs below the median of OOS Sharpe ratios. First, we need to calibrate the parameters of the Gumbel distribution associated with a set of Gaussian random variables. Suppose a random variable $max_N = max(\{SR_n | n = 1, \dots, N-1\})$, where SR_n is the Sharpe ratio estimated through a backtest for trial *n*. We know that the Gaussian distribution belongs to the Maximum Domain of Attraction of the Gumbel distribution, thus $max_N \sim \Lambda[\alpha, \beta]$, where α, β are the normalizing constants and Λ is the CDF of the Gumbel distribution. Next, we derive the values of these normalizing constants. It is known that the mean and standard deviation of a Gumbel distribution are

$$\begin{gathered} \mathrm{E}[max_N] = \alpha + \gamma\beta \\ \sigma[max_N] = \frac{\beta\pi}{\sqrt{6}} \end{gathered} \tag{4}$$

where γ is the Euler-Mascheroni constant. Applying the method of moments, we can derive the following:

- Given an estimate of $\hat{\sigma}[max_N]$, we can estimate $\hat{\beta} = \frac{\hat{\sigma}[max_N]\sqrt{6}}{\pi}$.
- Given an estimate of $\hat{\mathrm{E}}[max_N]$, and the previously obtained $\hat{\beta}$, we can estimate $\hat{\alpha} = \hat{\mathrm{E}}[max_N] - \gamma\hat{\beta}$.

These parameters allow us to model the distribution of the maximum Sharpe ratio IS out of a set of *N-1* trials. PBO can then be directly computed as $\phi = \phi_1 + \phi_2$, where:

$$\phi_1 = \int_{-\infty}^{2\widetilde{SR}} N\left[SR, \widetilde{SR}, \frac{1 + \frac{1}{2}\widetilde{SR}^2}{T}\right] (1 - \Lambda[max(0, SR), \alpha, \beta]) dSR \tag{5}$$

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

Probability ϕ_1 accounts for selecting IS a strategy with $SR_n = 0$, as a result of $SR_{N,IS} < SR_{IS}^*$. As we argued earlier, in this situation $SR_{OOS}^* \approx -max_N < Me[SR_{OOS}] = 0\ \ iif$ $SR_{IS}^* > 0$, hence the $max(0, SR)$ used to evaluate the Gumbel distribution. The integral has an upper boundary in $2\widetilde{SR}$ because beyond that point all trials lead to $SR_{OOS}^* < Me[SR_{OOS}]$, including the *N*th trial. That probability is accounted for by ϕ_2, which has a lower boundary of integration in $2\widetilde{SR}$. Snippet 2 implements the numerical integration of Eq. (16). As we will see in Section 7.3, the EVT estimates of PBO derived from Eq. (16) are in agreement with the MC estimates.

```
#!/usr/bin/env python
# On 20130704 by lopezdeprado@lbl.gov
#------------------------------------------------------------------------------------------
def testAccuracy_EVT(sr_base,sr_case):
    # Test accuracy by numerical integration
    # It does the same as testAccuracy_MC, but through numerical integration ...
    # ... of the base and case distributions.
    #1) Parameters
    parts,length,freq,minX,trials=1e4,1000,365.25*5/7.,-10,100
    emc=0.57721566490153286 # Euler-Mascheroni constant
    #2) SR distributions
    dist_base=[sr_base,((freq+.5*sr_base**2)/length)**.5]
    dist_case=(sr_case,((freq+.5*sr_case**2)/length)**.5)
    #3) Fit Gumbel (method of moments)
    maxList=[]
    for x in range(int(parts)):
            max_=max([gauss(dist_base[0],dist_base[1]) for i in range(trials)])
            maxList.append(max_)
    dist_base[1]=np.std(maxList)*6**.5/math.pi
    dist_base[0]=np.mean(maxList)-emc*dist_base[1]
    #4) Integration
    prob1=0
    for x in np.linspace(minX*dist_case[1],2*dist_case[0]-sr_base,parts):
            f_x=ss.norm.pdf(x,dist_case[0],dist_case[1])
            F_y=1-ss.gumbel_r.cdf(x,dist_base[0],dist_base[1])
            prob1+=f_x*F_y
    prob1*=(2*dist_case[0]-sr_base-minX*dist_case[1])/parts
    prob2=1-ss.norm.cdf(2*dist_case[0]-sr_base,dist_case[0],dist_case[1])
    print dist_base,dist_case
    print prob1,prob2,prob1+prob2
    return
```

Snippet 2 – Python code for computing PBO by EVT

7.3. EMPIRICAL STUDY OF ACCURACY

We are finally ready to evaluate the accuracy of CSCV's PBO. To achieve that, we will compare CSCV's PBO estimates against the two alternative benchmarks described in Sections 7.1 (MC) and 7.2 (EVT). Table 2 reports the results for a wide range of combinations of $\widetilde{SR}$ (SR_Case), T and N. Without loss of generality, $SR_n = 0$ for n=*1,...,N-1*. We do not need to consider alternative values of SR_n, because the likelihood of selecting the wrong strategy is a function of $\widetilde{SR} - SR_n$, not the absolute level of SR_n.

[TABLE 2 HERE]

Monte Carlo results were computed on 1,000 experiments. The proportion of IS optimal selections that underperformed OOS is reported in Prob_MC. Column Prob_EVT reports the corresponding PBO estimates, derived from Eq. (16). Because these latter results are derived from the actual distribution of the maximum SR, they are more accurate than the Monte Carlo estimates. In any case, EVT and MC results are very close, with a maximum absolute deviation of 4.2%.

We have computed CSCV's PBO on 1,000 randomly generated matrices $\boldsymbol{M}$ for every parameter combination $\left(\widetilde{SR}, T, N\right)$. Therefore, we have obtained 1,000 independent estimates of PBO for every parameter combination, with a mean and standard deviation reported in columns Mean_CSCV and Std_CSCV. This is not to be mistaken with the Monte Carlo result, which produced a single estimate of PBO out of 1,000 randomly generated matrices $\boldsymbol{M}$.

A comparison of the Mean_CSCV probability with the EVT result gives us an average absolute error of 2.1%, with a standard deviation of 2.9%. The maximum absolute error is 9.9%. That occurred for the combination $\left(\widetilde{SR}, T, N\right) = (3{,}500{,}500)$, whereby CSCV gave a more conservative estimate (24.7% instead of 14.8%). There is only one case where CSCV underestimated PBO, with an absolute error of 0.1%. The median error is only 0.7%, with a 5%-tile of 0% and a 95%-tile of 8.51%.

In conclusion, CSCV provides accurate estimates of PBO, with relatively small errors on the conservative side.

8. A PRACTICAL APPLICATION

Bailey et al. [2013] present an example of an investment strategy that attempts to profit from a seasonal effect. For the reader's convenience, we reiterate here how the strategy works. Suppose that we would like to identify the optimal monthly trading rule, given four customary parameters: Entry_day, Holding_period, Stop_loss and Side. *Side* defines whether we will hold long or short positions on a monthly basis. *Entry_day* determines the business day of the month when we enter a position. *Holding_period* gives the number of days that the position is held. *Stop_loss* determines the size of the loss as a multiple of the series' volatility that triggers an exit for that month's position. For example, we could explore all nodes that span the interval [1, …, 22] for *Entry_day*, the interval [1, …, 20] for *Holding_period*, the interval [0, …, 10] for *Stop_loss*, and [-1, 1] for *Sign*. The parameters combinations involved form a four-dimensional mesh of 8,800 elements. The optimal parameter combination can be discovered by computing the performance derived by each node.

As discussed in the above cited paper, a time series of 1000 daily prices (about 4 years) was generated by drawing from a random walk. Parameters were optimized (Entry_day = 11, Holding_period = 4, Stop_loss = -1 and Side = 1), resulting in an annualized Sharpe ratio is 1.27. Given the elevated Sharpe ratio, we could conclude that this strategy's performance is significantly greater than zero for any confidence level. Indeed, the PSR-Stat is 2.83, which implies a less than 1% probability that the true Sharpe ratio is below 0 (see Bailey and López de Prado [2012] for details).

We have estimated the PBO using our CSCV procedure, and obtained the results in Figure 6. Figure 6-1 shows that approx. 53% of the SR OOS are negative, despite all SR IS being positive and ranging between 1 and 2.2. Figure 6-2 plots the distribution of logits, which implies that, despite the elevated SR IS, the PBO is as high as 55%. Consequently, Figure 6-3 shows that the distribution of optimized OOS SR does not dominate the overall distribution of OOS SR. This is consistent with the fact that the underlying series follows a random walk, thus the serial independence among observations makes any seasonal patterns coincidental. The CSCV framework has succeeded in recognizing that the backtest was overfit.

[FIGURE 6 HERE]

Second, we generated a time series of 1000 daily prices (about 4 years), following a random walk. But unlike in the first case, we have shifted the returns of the first 5 random observations of each month to be centered at a quarter of a standard deviation. This generates a monthly seasonal effect, which the strategy selection procedure should discover. Figure 7 plots the random series, as well as the performance associated with the optimal parameter combination: Entry_day = 1, Holding_period = 4, Stop_loss = -10 and Side = 1. The annualized Sharpe ratio is 1.54.

[FIGURE 7 HERE]

Figure 8 reports the results of the CSCV analysis, which confirm the validity of this backtest in the sense that performance inflation from overfitting is minimal. Figure 8-1 shows that only 13% of the OOS SR to be negative. Because there is a real monthly effect in the data, the PBO for this second case should be substantially lower than the PBO of the first case. Figure 8-2 shows a distribution of logits with a PBO of only 13%. Figure 8-3 evidences that the distribution of OOS SR from IS optimal combinations clearly dominates the overall distribution of OOS SR.

[FIGURE 8 HERE]

In this practical application we have illustrated how simple is to produce overfit backtests when answering common investment questions, such as the presence of seasonal effects. We refer the reader to Appendix 4 for the implementation of this experiment in Python language. Similar experiments can be designed to demonstrate overfitting in the context of other effects, such as trend-following, momentum, mean-reversion, event-driven effects, etc. Given the facility with which elevated Sharpe ratios can be manufactured IS, the reader would be well advised to remain critical of backtests and researchers that fail to report the PBO.

9. CONCLUSIONS

Bailey and López de Prado [2012] developed methodologies to evaluate the probability that a Sharpe ratio is inflated (PSR), and the minimum track record length (MinTRL) required for a Sharpe ratio to be statistically significant. These statistics were developed to assess Sharpe ratios based on live performance (track record). We have not been able to find similar statistics or methodologies applicable to evaluate backtested Sharpe ratios. The representativeness of backtested performance estimates has been the subject of this study.

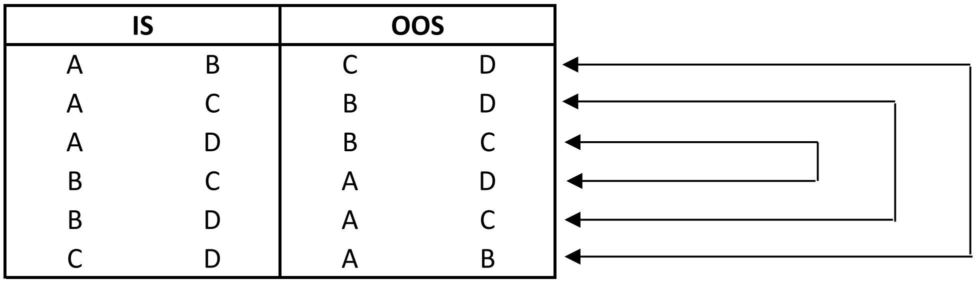
Standard statistical techniques designed to detect overfitting in the context of regression models are poorly equipped to assess backtest overfitting. Hold-outs in particular are unreliable and easy to manipulate. As long as a researcher tries more than one strategy configuration, overfitting is always present. However, hold-out methods do not take into account the number of trials involved in the strategy selection (N), making it an inappropriate method to evaluate a backtest's representativeness.

The procedure presented in this paper has specifically been designed to determine the *probability of backtest overfitting* (PBO). This is defined as the probability that the strategy with optimal performance IS delivers OOS a performance below the median performance of all trials attempted by the researcher. When that probability is high, optimizing IS has a detrimental effect in terms of OOS performance, because the backtest profits from specific features in the IS subset that are not present elsewhere. CSCV identifies such situation by generating a large number of combinations of IS subsets, and determining the proportion of those where overfitting has taken place. Unlike hold-out tests, CSCV takes into account the number of trials attempted (N). Unlike in the case of hold-out tests, the decision as to whether overfitting occurs does not depend on an arbitrary division of the data into an IS and an OOS set, but on all possible combinations. In addition, all IS subsets are re-used as OOS, and all OOS subsets are re-used as IS.

We have assessed the accuracy of CSCV with regards to the estimation of PBO in two different ways, on a wide variety of test cases. Monte Carlo simulations show that CSCV applied on a single dataset provides similar results to computing PBO on a large number of independent samples. We have also computed directly PBO by deriving the Extreme Value distributions that model the performance of IS optimal strategies. Results indicate that CSCV provides accurate estimates of PBO, with relatively small errors on the conservative side.

In conclusion, we believe that CSCV is a new and powerful tool in the arsenal of investors and financial markets' researchers. At least we hope that this study raises the awareness concerning the futility of computing and reporting backtest results without controlling for its PBO and MinBTL.

TABLES



IS		OOS	
A	B	C	D
A	C	B	D
A	D	B	C
B	C	A	D
B	D	A	C
C	D	A	B

Table 1 – Generating the C_S symmetric combinations

Table 1 shows the six combinations of four subsamples A, B, C, D, grouped in two subsets of size two. The first subset is the training set (or in-sample). This is used to determine the optimal model configuration. The second subset is the testing set (or out-of-sample), on which the in-sample optimal model configuration is tested. Running the *N* model configurations over each of these combinations allows us to derive a relative ranking, expressed as a logit. The outcome is a distribution of logits, one per combination. Note that each training subset combination is re-used as a testing subset and viceversa.

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

Table 2 – CSCV's accuracy

Table 2 shows PBO estimates using three alternative methods: Combinatorially Symmetric Cross-Validation (CSCV), Monte Carlo (MC) and Extreme Value Theory (EVT).

First, we have computed CSCV's PBO on 1,000 randomly generated matrices M for every parameter combination $\left(\widetilde{SR},T,N\right)$. This has provided us with 1,000 independent estimates of PBO for every parameter combination, with a mean and standard deviation reported in columns Mean_CSCV and Std_CSCV. Second, we generated 1,000 matrices M (experiments) for various test cases of order (TxN)=(1000x100), and computed the proportion of experiments that yielded an OOS performance below the median. The proportion of IS optimal selections that underperformed OOS is reported in Prob_MC. This Prob_MC is well within the confidence bands implied by Mean_CSCV and Std_CSCV.

A comparison of the Mean_CSCV probability with the EVT result gives us an average absolute error is 2.1%, with a standard deviation of 2.9%. The maximum absolute error is 9.9%. That occurred for the combination $\left(\widetilde{SR},T,N\right)$=(3,500,500), whereby CSCV gave a more conservative estimate (24.7% instead of 14.8%). There is only one case where CSCV underestimated PBO, with an absolute error of 0.1%. The median error is only 0.7%, with a 5%-tile of 0% and a 95%-tile of 8.51%.

In conclusion, CSCV provides accurate estimates of PBO, with relatively small errors on the conservative side.

FIGURES

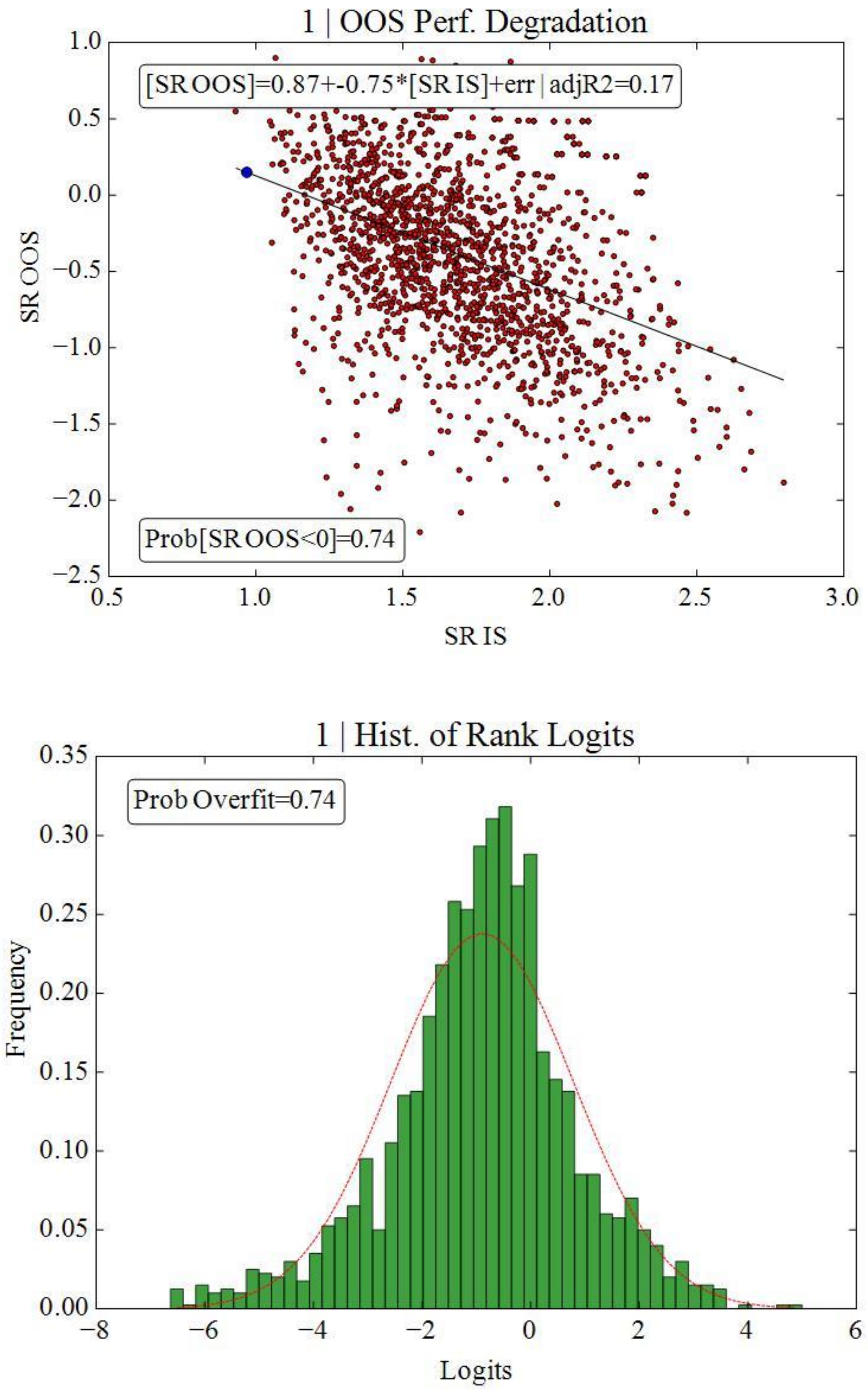


Figure 1 – Performance degradation and distribution of logits

The CSCV framework provides five analysis of backtest overfitting: i) Out-Of-Sample Performance Degradation, ii) Out-Of-Sample Probability of Loss, iii) Probability of Overfitting (PBO), and iv) Backtest Stochastic Dominance. Figure 1 provides a graphical representation of the first three. The upper plot shows that pairs of (SR IS, SR OOS) for the optimal model configurations selected for each subset $c \in C_S$, which corresponds to analysis i). This also allows us to compute the proportion of combinations with negative performance, $Prob\left[\left[\overline{R_{n^*}}\right]_c < 0\right]$, which corresponds to analysis ii). The lower plot displays the distribution of logits, which allows us to compute the probability of backtest overfitting (PBO), or analysis iii). This represents the rate at which optimal IS strategies underperform the median of the OOS trials.

Note that, even if $\phi \approx 0$, $Prob\left[\left[\overline{R_{n^*}}\right]_c < 0\right]$ could be high, in which case the strategy's performance OOS is poor for reasons other than overfitting.

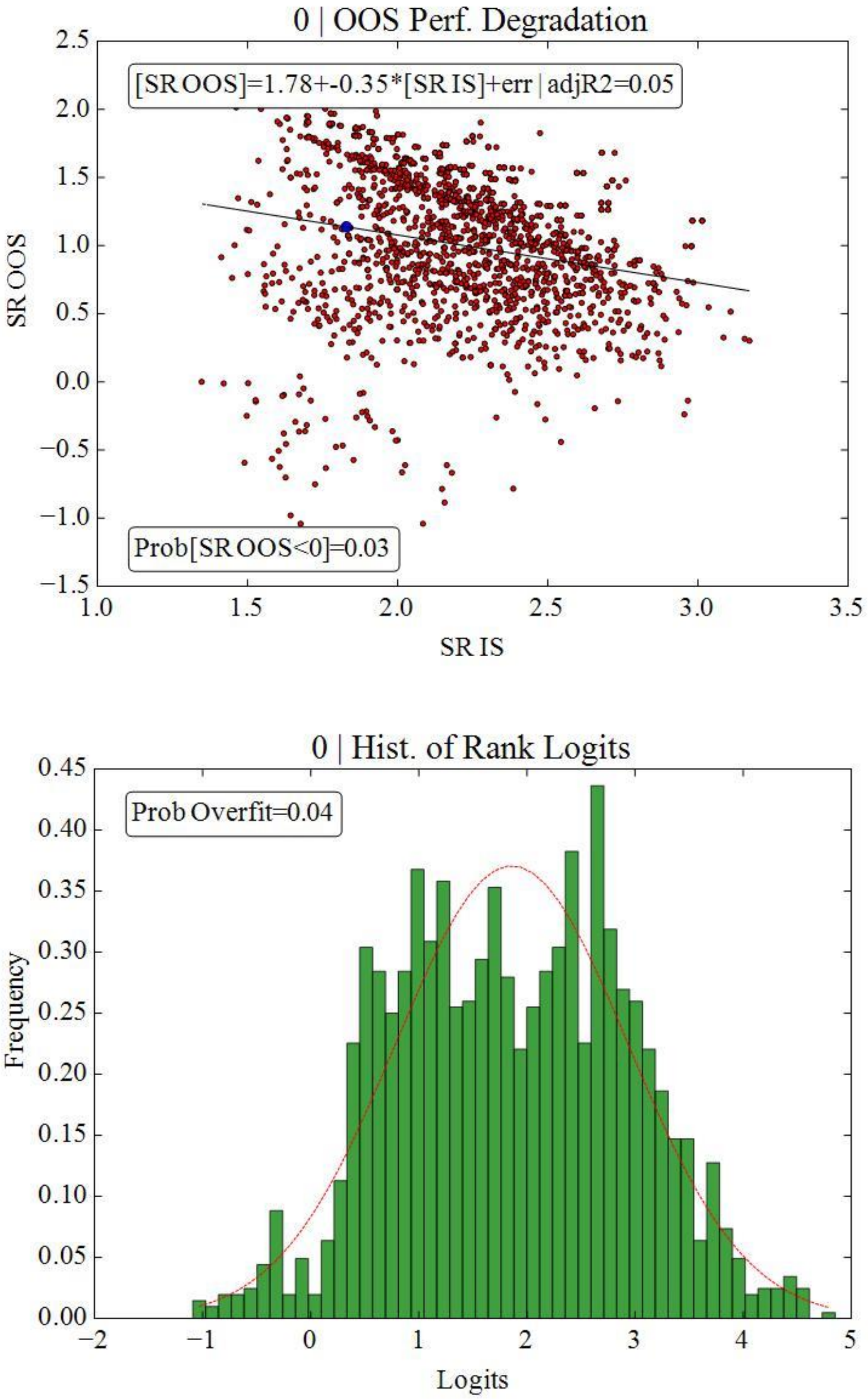


Figure 2 – Performance degradation and distribution of logits for a real investment strategy

Figure 2 provides the performance degradation and distribution of logits of a real investment strategy. Unlike in the previous example, the OOS probability of loss is very small (about 3%), and the proportion of selected (IS) model configurations that performed OOS below the median of overall model configurations was only 4%.

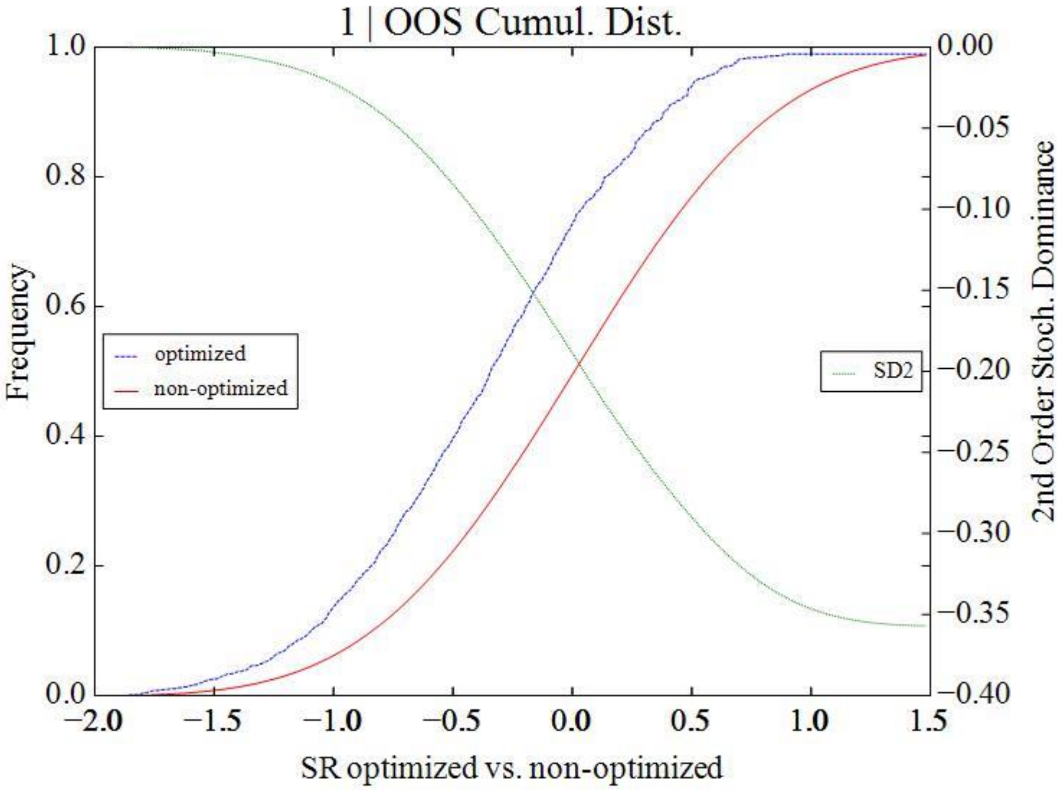


Figure 3 – Stochastic dominance (example 1)

Figure 3 complements the analyses presented in Figure 1, with the 4th analysis listed in Section 3. Stochastic dominance allows us to rank gambles or lotteries without having to make strong assumptions regarding an individual's utility function. In the context of our framework, first-order stochastic dominance occurs if $Prob\left[\overline{R_{n^*}} \geq x\right] \geq Prob\left[\overline{\boldsymbol{R}} \geq x\right]$ for all x, and for some x, $Prob\left[\overline{R_{n^*}} \geq x\right] > Prob\left[\overline{\boldsymbol{R}} \geq x\right]$. A less demanding criterion is second-order stochastic dominance: $SD2[x] = \int_{-\infty}^{x}\left(Prob\left[\overline{\boldsymbol{R}} \leq x\right] - Prob\left[\overline{R_{n^*}} \leq x\right]\right)dx \geq 0$ for all x, and that $SD2[x] > 0$ at some x.

In this example, the distribution of OOS SR of optimized (IS) model configurations does not dominate (in first order) the distribution of OOS SR of overall model configurations. This can be appreciated in the fact that for every level of OOS SR the proportion of optimized model configurations is greater than the proportion of non-optimized, thus the probabilistic mass of the former is shifted to the left of the non-optimized. SD2 plots the second order stochastic dominance, which indicates that the distribution of optimized model configurations does not dominate the non-optimized according to this less demanding criterion.

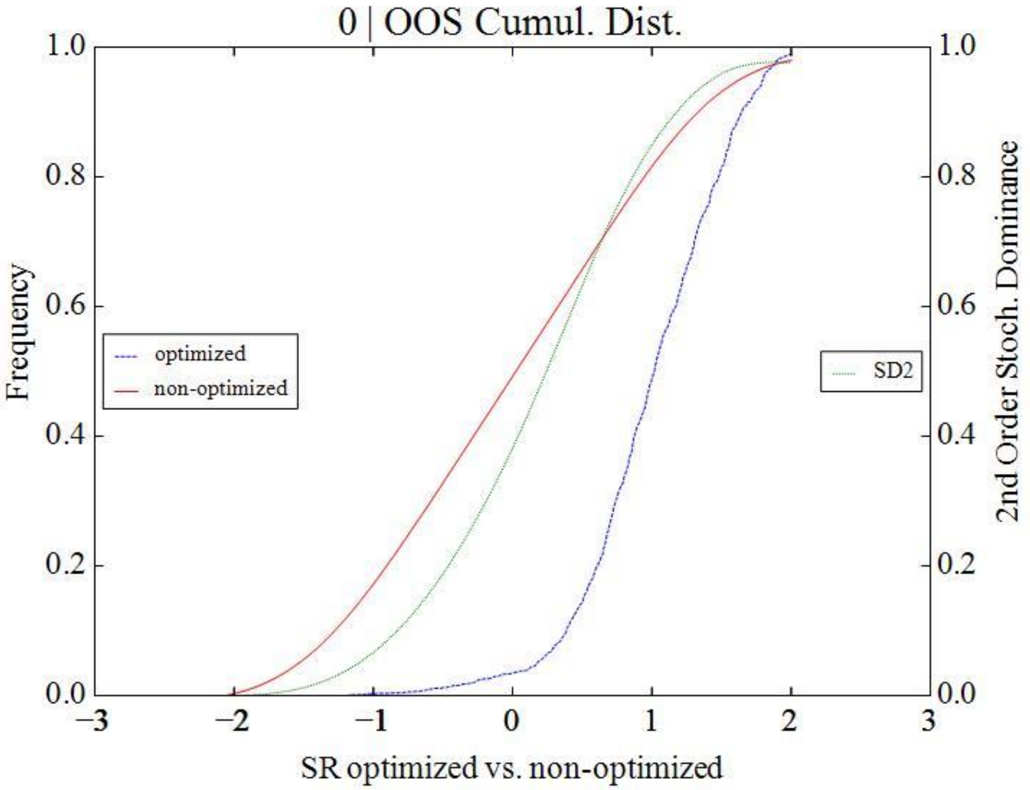


Figure 4 – Stochastic dominance (example 2)

In the example displayed by Figure 4, the distribution of OOS SR of optimized (IS) model configurations dominates the distribution of OOS SR of overall model configurations. For every level of OOS SR the proportion of optimized model configurations is lower than the proportion of non-optimized, thus the probabilistic mass of the former is shifted to the right of the non-optimized. First order stochastic dominance implies second order stochastic dominance, and consequently the SD2 function is positive for all OOS SR.

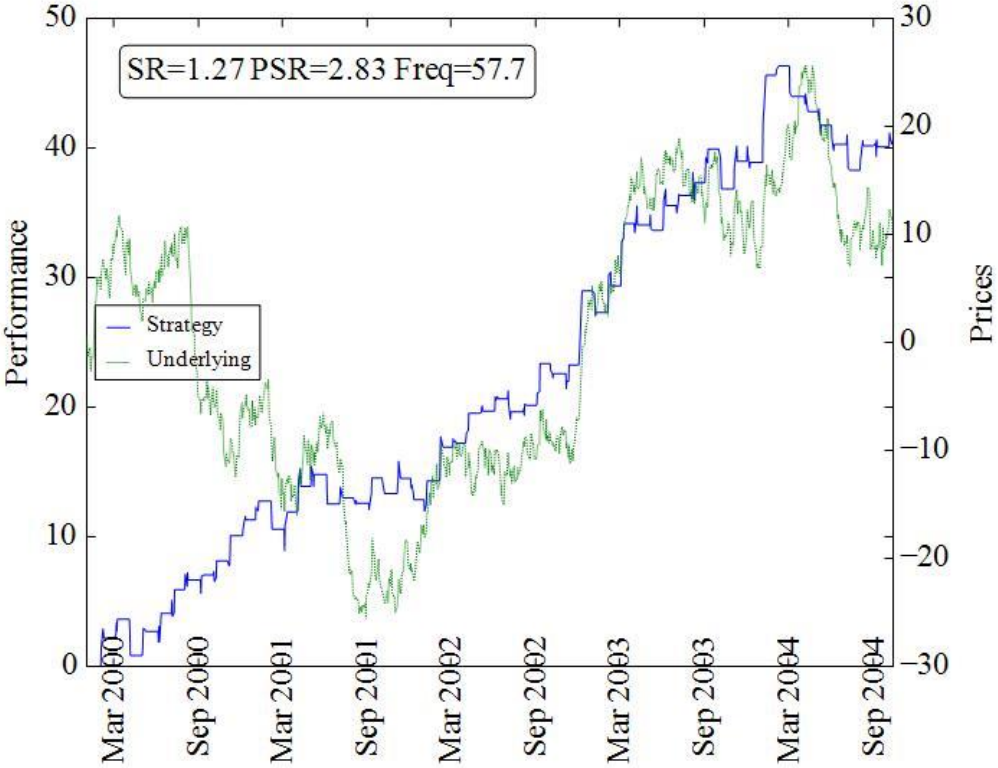
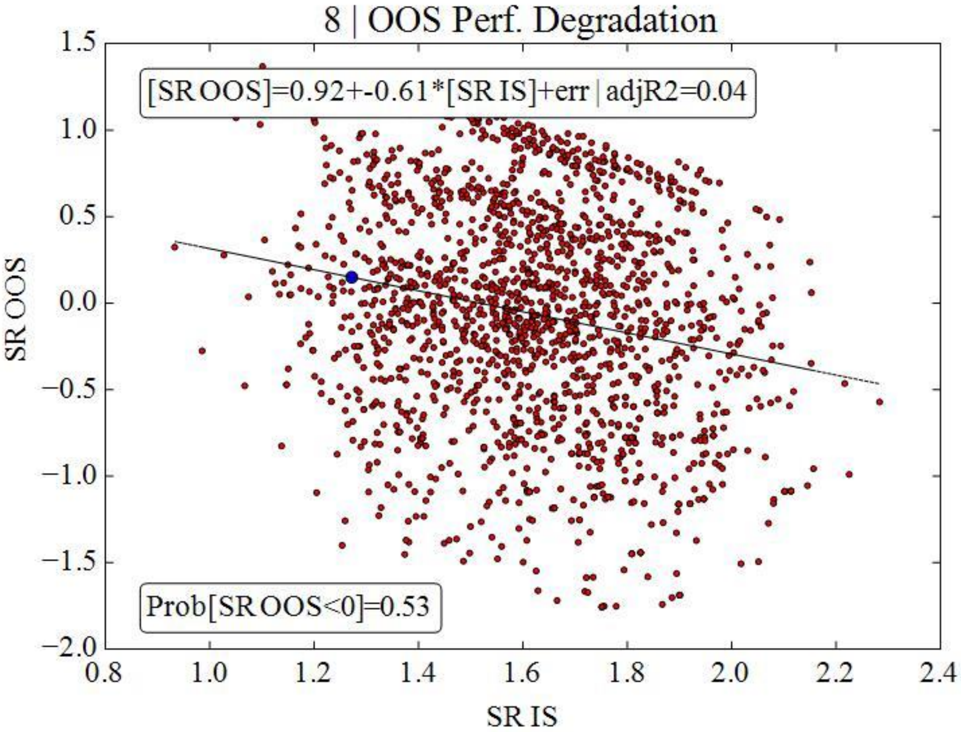
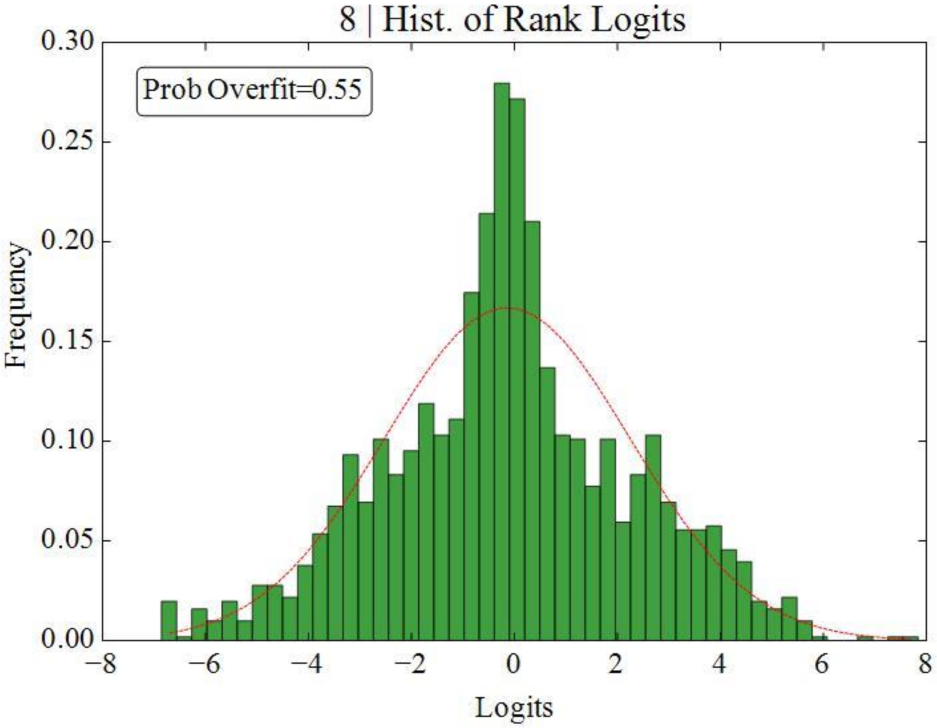


Figure 5 – Backtested performance of a seasonal strategy (example 1)

We have generated a time series of 1000 daily prices (about 4 years), following a random walk. The PSR-Stat of the optimal model configuration is 2.83, which implies a less than 1% probability that the true Sharpe ratio is below 0. Consequently, we have been able to identify a seasonal strategy with a SR of 1.27 despite the fact that no seasonal effect exists.


8 | OOS Perf. Degradation
[SR OOS]=0.92+-0.61*[SR IS]+err | adjR2=0.04
Prob[SR OOS<0]=0.53
SR OOS
SR IS
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4


8 | Hist. of Rank Logits
Prob Overfit=0.55
Frequency
Logits
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−8
−6
−4
−2
0
2
4
6
8

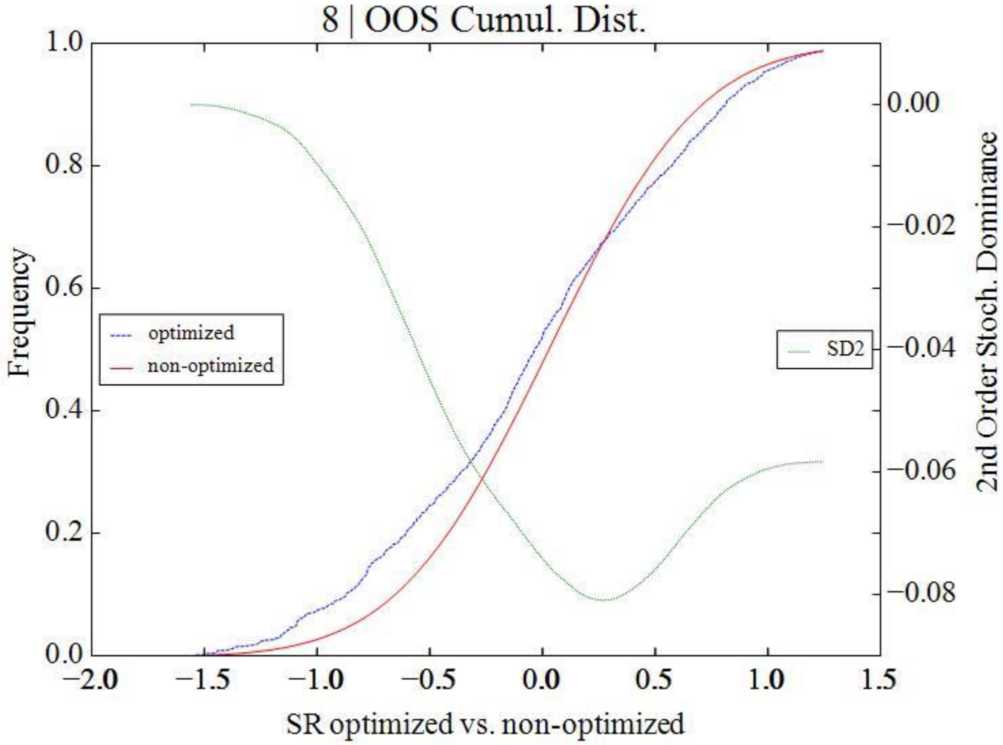


Figure 6 – CSCV analysis of the backtest of a seasonal strategy (example 1)

OOS SR of optimal configurations is negative in 53% of cases. The distribution of logits implies that, despite the elevated SR IS, the PBO is as high as 55%. Consequently, the distribution of optimized OOS SR does not dominate the overall distribution of OOS SR. The CSCV analysis has succeeded in rejecting the overfit backtest.

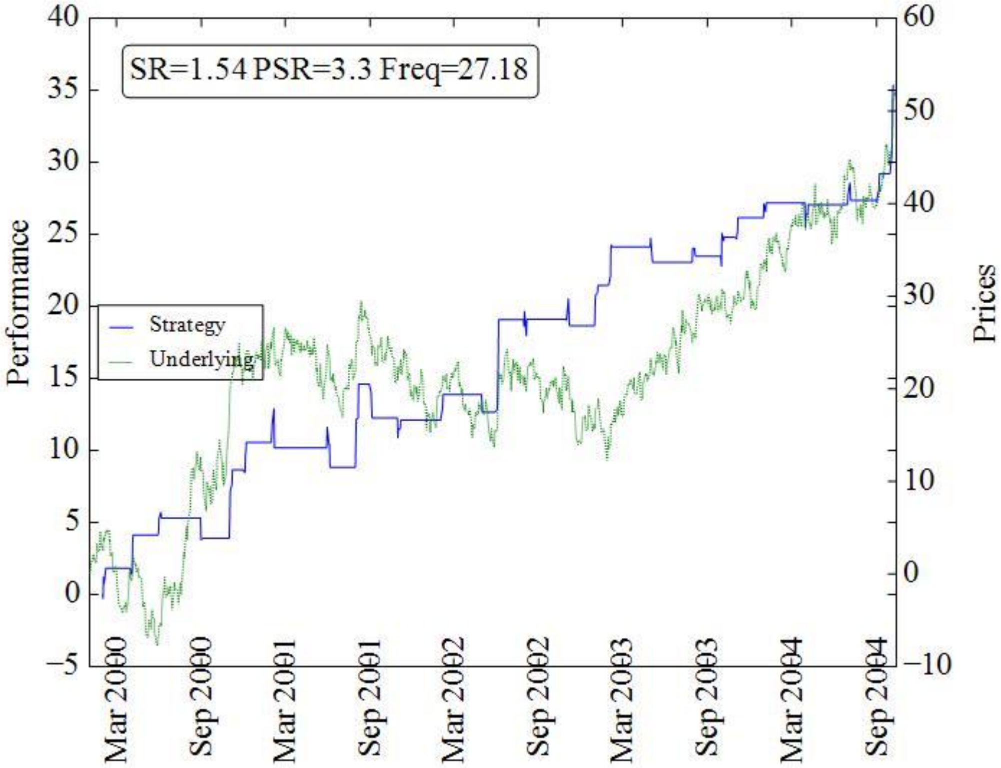
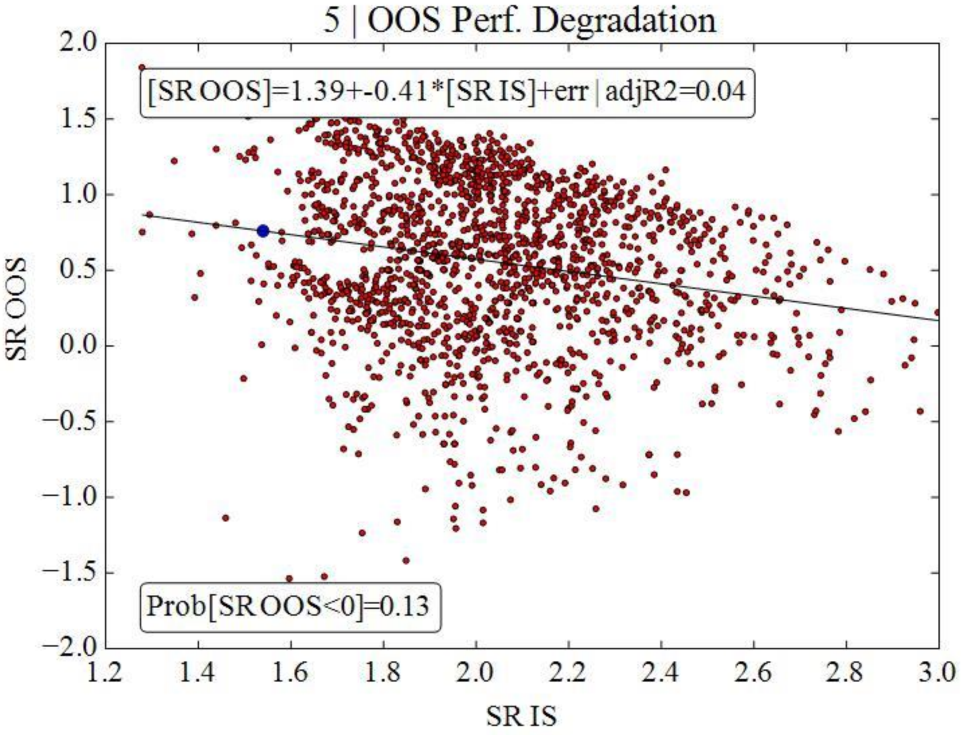
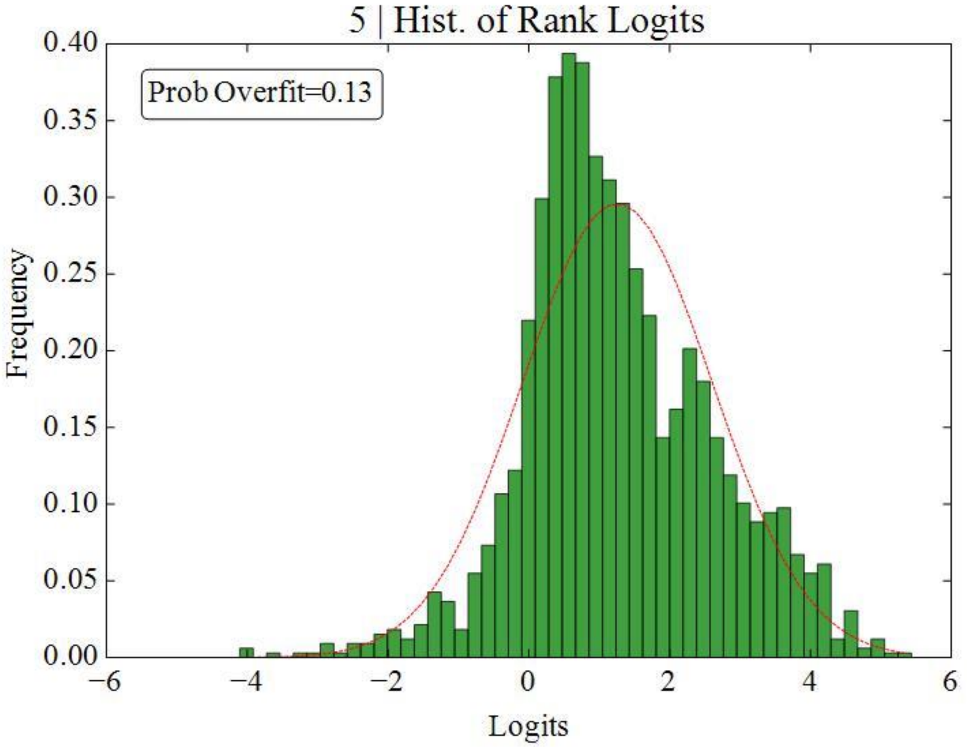


Figure 7 – Backtested performance of a seasonal strategy (example 2)

We have taken the previous 1000 series and shifted the returns of the first 5 observations of each month by a quarter of a standard deviation. This generates a monthly seasonal effect, which our strategy selection procedure should discover. The Sharpe Ratio is similar to the previous (overfit) case (1.5 vs. 1.3).


5 | OOS Perf. Degradation
[SR OOS]=1.39+-0.41*[SR IS]+err | adjR2=0.04
Prob[SR OOS<0]=0.13
SR OOS
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
SR IS


5 | Hist. of Rank Logits
Prob Overfit=0.13
Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6
Logits

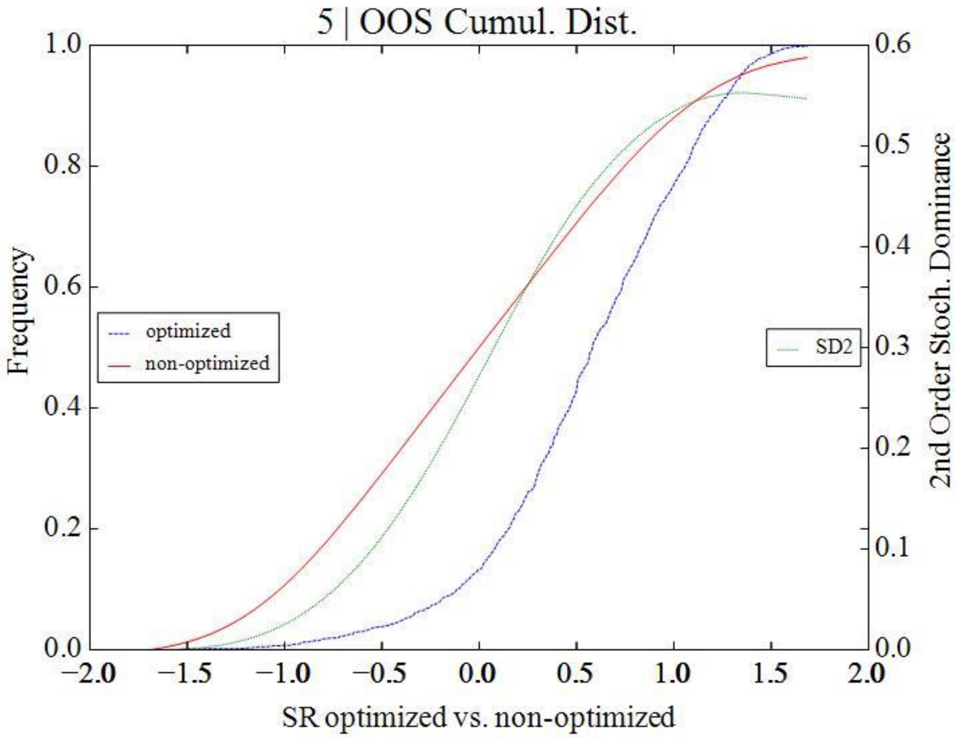


Figure 8 – CSCV analysis of the backtest of a seasonal strategy (example 2)

The SR OOS of optimal configurations is negative in only 13% of cases (compared to 53%). The distribution of logits implies that the PBO is only 13%. Consistently, the distribution of optimized OOS SR dominates (in first and second order) the overall distribution of OOS SR. The CSCV analysis has correctly recognized the validity of this backtest, in the sense that performance inflation from overfitting is small.

REFERENCES

- Bailey, D., J. Borwein, M. López de Prado and J. Zhu (2013): "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-Of-Sample Performance" Working paper. Available at http://ssrn.com/abstract=2308659
- Bailey, D. and M. López de Prado (2012): "The Sharpe Ratio Efficient Frontier," *Journal of Risk*, 15(2), pp. 3-44. Available at http://ssrn.com/abstract=1821643.
- Embrechts, P., C. Klueppelberg and T. Mikosch (2003): "Modelling Extremal Events," Springer Verlag, New York.
- Feynman, R. (1964): "The Character of Physical Law," The MIT Press.
- Hadar, J. and W. Russell (1969): "Rules for Ordering Uncertain Prospects," *American Economic Review*, Vol. 59, pp. 25-34.
- Harvey, C. and Y. Liu (2013): "Backtesting", SSRN, working paper.
- Hawkins, D. (2004): "The problem of overfitting," *Journal of Chemical Information and Computer Science*, Vol. 44, pp. 1-12.
- Hirsch, Y. (1987): "Don't Sell Stocks on Monday," Penguin Books, 1st Edition.
- Leinweber, D. and K. Sisk (2011): "Event Driven Trading and the 'New News'," *Journal of Portfolio Management*, Vol. 38(1), 110-124.
- Lo, A. (2002): "The Statistics of Sharpe Ratios," *Financial Analysts Journal*, (58)4, July/August.
- López de Prado, M. and A. Peijan (2004): "Measuring the Loss Potential of Hedge Fund Strategies," *Journal of Alternative Investments*, Vol. 7(1), pp. 7-31. Available at http://ssrn.com/abstract=641702.
- López de Prado, M. and M. Foreman (2012): "A Mixture of Gaussians approach to Mathematical Portfolio Oversight: The EF3M algorithm," working paper, RCC at Harvard University. Available at http://ssrn.com/abstract=1931734.
- Resnick, S. (1987): "Extreme Values, Regular Variation and Point Processes," Springer.
- Romano, J. and M. Wolf (2005): "Stepwise multiple testing as formalized data snooping", Econometrica 73(4), pp. 1273-1282.
- Schorfheide, F. and K. Wolpin (2012): "On the Use of Holdout Samples for Model Selection," *American Economic Review*, 102(3), pp. 477-481.
- Van Belle, G. and K. Kerr (2012): "Design and Analysis of Experiments in the Health Sciences," John Wiley & Sons.
- Weiss, S. and C. Kulikowski (1990): "Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems," Morgan Kaufman, 1st Edition.
- White, H. (2000): "A Reality Check for Data Snooping", *Econometrica*, 68(5), pp. 1097-1126.
- Wittgenstein, Ludwig (1953): "Philosophical Investigations," Prentice Hall.