THE PROBABILITY OF BACKTEST OVERFITTING

David H. Bailey [*] Jonathan M. Borwein [†]

Marcos López de Prado [‡] Qiji Jim Zhu[§]

February 27, 2015

Revised version: February 2015

[*] Lawrence Berkeley National Laboratory (retired), 1 Cyclotron Road, Berkeley, CA 94720, USA, and Research Fellow at the University of California, Davis, Department of Computer Science. E-mail: `david@davidhbailey.com`; URL: `http://www.davidhbailey.com`

[†] Laureate Professor of Mathematics at University of Newcastle, Callaghan NSW 2308, Australia, and a Fellow of the Royal Society of Canada, the Australian Academy of Science, the American Mathematical Society and the AAAS. E-mail: `jonathan.borwein@newcastle.edu.au`; URL: `http://www.carma.newcastle.edu.au/jon`

[‡] Senior Managing Director at Guggenheim Partners, New York, NY 10017, and Research Affiliate at Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. E-mail: `lopezdeprado@lbl.gov`; URL: `http://www.QuantResearch.info`

[§] Professor, Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008, USA. Email: `zhu@wmich.edu`; URL: `http://homepages.wmich.edu/~zhu/`

THE PROBABILITY OF BACKTEST OVERFITTING

Abstract

Many investment firms and portfolio managers rely on backtests (i.e., simulations of performance based on historical market data) to select investment strategies and allocate 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 general framework to assess the probability of backtest overfitting (PBO). We illustrate this framework with specific generic, model-free and nonparametric implementations in the context of investment simulations, which implementations we call combinatorially symmetric cross-validation (CSCV). We show that CSCV produces reasonable estimates of PBO for several useful examples.

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.

Acknowledgements. We are indebted to the Editor and three anonymous referees who peer-reviewed this article as well as a related article for the *Notices of the American Mathematical Society* [1]. We are also grateful to Tony Anagnostakis (Moore Capital), Marco Avellaneda (Courant Institute, NYU), Peter Carr (Morgan Stanley, NYU), Paul Embrechts (ETH Zürich), Matthew D. Foreman (University of California, Irvine), Ross Garon (SAC Capital), Jeffrey S. Lange (Guggenheim Partners), Attilio Meucci (KKR, NYU), Natalia Nolde (University of British Columbia and ETH Zürich) and Riccardo Rebonato (PIMCO, University of Oxford) for many useful and stimulating exchanges.

Sponsorship. Research 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 and by various Australian Research Council grants.

> "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 [36].

1 Introduction

Modern investment strategies rely on the discovery of patterns that can be quantified and monetized in a systematic way. For example, algorithms can be designed to profit from phenomena such as "momentum," i.e., the tendency of many securities to exhibit long runs of profits or losses, beyond what could be expected from securities following a martingale. One advantage of this systematization of investment strategies is that those algorithms are amenable to "backtesting." A backtest is a historical simulation of how an algorithmic strategy would have performed in the past. Backtests are valuable tools because they allow researchers to evaluate the risk/reward profile of an investment strategy before committing funds.

Recent advances in algorithmic research and high-performance computing have made it nearly trivial to test millions and billions of alternative investment strategies on a finite dataset of financial time series. While these advances are undoubtedly useful, they also present a negative and often silenced side-effect: The alarming rise of false positives in related academic publications (The Economist [32]). This paper introduces a computational procedure for detecting false positives in the context of investment strategy research.

To motivate our study, consider a researcher who is investigating an algorithm to profit from momentum. Perhaps the most popular technique among Commodity Trading Advisors (CTAs) is to use so-called crossing-moving averages to detect a change of trend in a security[1]. Even for the simplest case, there are at least five parameters that the researcher can fit: Two sample lengths for the moving averages, entry threshold, exit threshold and stop-loss. The number of combinations that can be tested over thousands of securities is in the billions. For each of those billions of backtests, we could estimate its Sharpe ratio (or any other performance statistic), and determine whether that Sharpe ratio is indeed statistically significant at a confidence level of 95%. Although this approach is consistent with the Neyman-Pearson framework of hypothesis testing, it is highly likely that false positives will emerge with a probability greater than 5%. The reason

[1] Several technical tools are based on this principle, such as the Moving Average Convergence Divergence (MACD) indicator.

is that a 5% false positive probability only holds when we apply the test exactly once. However, we are applying the test on the same data multiple times (indeed, billions of times), making the emergence of false positives almost certain.

The core question we are asking is this: *What constitutes a legitimate empirical finding in the context of investment research?* This may appear to be a rather philosophical question, but it has important practical implications, as we shall see later in our discussion. Financial discoveries typically involve identifying a phenomenon with low signal-to-noise ratio, where that ratio is driven down as a result of competition. Because the signal is weak, a test of hypothesis must be conducted on a large sample as a way of assessing the existence of a phenomenon. This is not the typical case in scientific areas where the signal-to-noise ratio is high. By way of example, consider the apparatus of classical mechanics, which was developed centuries before Neyman and Pearson proposed their theory of hypothesis testing. Newton did not require statistical testing of his gravitation theory, because the signal from that phenomenon dominates the noise.

The question of 'legitimate empirical findings' is particularly troubling when researchers conduct multiple tests. The probability of finding false positives increases with the number of tests conducted on the same data (Miller [25]). As each researcher carries out millions of regressions (Sala-i-Martin [28]) on a finite number of independent datasets without controlling for the increased probability of false positives, some researchers have concluded that 'most published research findings are false' (see Ioannidis [17]).

Furthermore, it is common practice to use this computational power to *calibrate* the parameters of an investment strategy in order to maximize its performance. But because the signal-to-noise ratio is so weak, often the result of such calibration is that parameters are chosen to profit from past noise rather than future signal. The outcome is an overfit backtest [1]. Scientists at Lawrence Berkeley National Laboratory have developed an online tool to demonstrate this phenomenon. This tool generates a time series of pseudorandom returns, and then calibrates the parameters of an optimal monthly strategy (i.e., the sequence of days of the month to be long the security, and the sequence of days of the month to be short). After a few hundred iterations, it is trivial to find highly profitable strategies in-sample, despite the small number of parameters involved. Performance out-of-sample is, of course, utterly disappointing. The tool is available at `http://datagrid.lbl.gov/backtest/index.php`.

Backtests published in academic or practitioners' publications almost never declare the number of trials involved in a discovery. Because those researchers have most likely not controlled for the number of trials, it is highly probable that their findings constitute false positives ([1, 3]). Even though researchers at academic and investment institutions may be aware of these problems, they have little incentive to expose them. Whether their motivations are to receive tenure or raise funds for a new systematic fund, those researchers would rather ignore this problem and make their investors or managers believe that backtest overfitting does not affect their results. Some may even pretend that they are controlling for overfitting using inappropriate techniques, exploiting the ignorance of their sponsors, as we will see later on when discussing the 'hold-out' method.

The goal of our paper is to develop computational techniques to control for the increased probability of false positives as the number of trials increases, applied to the particular field of investment strategy research. For instance, journal editors and investors could demand researchers to estimate that probability when a backtest is submitted to them.

Our approach. First, we introduce a precise characterization of the event of backtest overfitting. The idea is simple and intuitive: For overfitting to occur, the strategy configuration that delivers maximum performance *in sample* (IS) must systematically underperform the remaining configurations *out of sample* (OOS). Typically the principal reason for this underperformance is that the IS "optimal" strategy is so closely tied to the noise contained in the training set that further optimization of the strategy becomes pointless or even detrimental for the purpose of extracting the signal.

Second, we establish a general framework for assessing the probability of the event of backtest overfitting. We model this phenomenon of backtest overfitting using an abstract probability space in which the sample space consist of pairs of IS and OOS test results.

Third, we set as null hypothesis that backtest overfitting has indeed taken place, and develop an algorithm that tests for this hypothesis. For a given strategy, the *probability of backtest overfitting* (PBO) is then evaluated as the conditional probability that this strategy underperforms the median OOS while remaining optimal IS. While the PBO provides a direct way to quantify the likelihood of backtest overfitting, the general framework also affords us information to look into the overfitting issue from different perspectives. For example, besides PBO, this framework can also be used to assess performance decay, probability of loss, and possible stochastic dominance of a strategy.

It is worth clarifying in what sense do we speak of a probability of backtest overfitting. Backtest overfitting is a deterministic fact (either the model is overfit or it is not), hence it may seem unnatural to associate a probability to a non-random event. Given some empirical evidence and priors, we can infer the posterior probability that overfitting has taken place. Examples of this line of reasoning abound in information theory and machine learning treatises, e.g. [23]. It is in this Bayesian sense that we define and estimate PBO.

A generic, model-free, and nonparametric testing algorithm is desirable, since backtests are applied to trading strategies produced using a great variety of different methods and models. For this reason, we present a specific implementation, which we call a *combinatorially symmetric cross-validation* (CSCV). We show that CSCV produces reasonable estimates of PBO for several useful examples.

Our CSCV implementation draws from elements in experimental mathematics, information theory, Bayesian inference, machine learning and decision theory to address the very particular problem of assessing the representativeness of a backtest. This is not an easy problem, as evidenced by the scarcity of academic papers addressing a dilemma that most investors face. This gap in the literature is disturbing, given the heavy reliance on backtests among practitioners. One advantage of our solution is that it only requires time series of backtested performance. We avoid the credibility issue of preserving a truly out-of-sample test-set by not requiring a fixed "hold-out," and swapping all in-sample (IS) and out-of-sample (OOS) datasets. Our approach is generic in the sense of not requiring knowledge of either the trading rule or forecasting equation. The output is a bootstrapped distribution of OOS performance measure. Although in our examples we measure performance using the Sharpe ratio, our methodology does not rely on this particular performance statistic, and it can be applied to any alternative preferred by the reader.

We emphasize that the CSCV implementation is only one illustrative technique. The general framework is flexible enough to accommodate other task-specific methods for estimating the PBO.

Comparisons to other approaches. Perhaps the most common approach to prevent overfitting among practitioners is to require the researcher to withhold a portion of the available data sample for separate testing and validation as OOS performance (this is known as the "hold-out" or "test set" method). If the IS and OOS performance levels are congruent, the investor might decide to "reject" the hypothesis that the backtest is overfit.

The main advantage of this procedure is its simplicity. This approach is, however, 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 dataset. Second, even if no "hold-out" data was used, any seasoned researcher knows well how financial variables performed over the time period covered by the OOS dataset, and that information may well be used in the strategy design, consciously or not (see Schorfheide and Wolpin [29]).

Third, hold-out is clearly inadequate for small samples—the IS dataset will be too short to fit, and the OOS dataset too short to conclude anything with sufficient confidence. Weiss and Kulikowski [34] argue that hold-out should not be applied to an analysis with less than $1,000$ observations. For example, if a strategy trades on a weekly basis, hold-out should not be used on backtests of less than 20 years. Along the same lines, Van Belle and Kerr [33] point out the high variance of hold-out estimation errors. If one is unlucky, the chosen hold-out section may be the one that refutes a valid strategy or supports an invalid strategy. Different hold-outs are thus likely to lead to different conclusions.

Fourth, even if the researcher works with a large sample, the OOS analysis will need to consume a large proportion of the sample to be conclusive, which is detrimental to the strategy's design (see Hawkins [15]). 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 has been done on arguably 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. [1] 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.

In short, the hold-out method 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 (input ignored by hold-out). In this paper we will present a way to compute this probability.

Another approach popular among practitioners consists in modeling the underlying financial variable by generating pseudorandom scenarios and measuring the performance of the resulting investment strategy for those scenarios (see Carr and López de Prado [6] for a valid application of this technique). This approach has 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, or may not contain all relevant statistical features, and may need to be customized to every variable (with large development costs). Some retail trading platforms offer backtesting procedures based on this approach, such as by pseudorandom generation of tick data by fractal interpolation.

Several procedures have been proposed to determine whether an econometric model is overfit. See White [35], Romano et al. [27], Harvey et al. [13] 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. [1]. Investment strategies in general are not amenable to characterization through a system of algebraic equations. Regression-tree decision making, for example, requires a hierarchy that only combinatorial frameworks like graph theory can provide, and which are beyond the geometric arguments used in econometric models (see Calkin and López de Prado [4, 5]). On the other hand, the approach proposed here shares the same philosophy in that both are trying to assess the probability of overfitting.

Structure of the paper. The rest of the study is organized as follows: Section 2 sets the foundations of our framework: we describe our general framework for the backtest overfitting probability in Subsection 2.1 and present the CSCV method for estimate this probability in Subsection 2.2. Section 3 discusses other ways that our general framework can be used to assess a backtest. Section 4 further discusses some of the features of the CSCV method, and how it relates to other machine learning methods. Section 5 lists some of the limitations of this method. Section 6 discusses a practical application, and Section 7 summarizes our conclusions. We have carried out several test cases to illustrate how the PBO compares to different scenarios, and to assess the accuracy of our method using two alternative approaches (Monte Carlo Methods and Extreme Value Theory). The interested reader can find the details of those studies following this link: `http://ssrn.com/abstract=2568435`

2 The framework

2.1 Definition of overfitting in the context of strategy selection

We first establish a measure theoretic framework in which the probability of backtest overfitting and other statistics related to the issue of overfitting can be rigorously defined. Consider a probability space $(\mathcal{T}, \mathcal{F}, Prob)$ where $\mathcal{T}$ represents a sample space of pairs of IS and OOS samples. We aim at estimating the probability of overfitting for the following *backtest strategy selection process*: select from N strategies labeled as $(1, 2, \ldots, N)$ the 'best' one using backtesting according to a given performance measure, say, the Sharpe ratio. Fixing a performance measure, we will use random vectors $\mathbf{R} = (R_1, R_2, \ldots, R_N)$ and $\overline{\mathbf{R}} = (\overline{R_1}, \overline{R_2}, \ldots, \overline{R_N})$ on $(\mathcal{T}, \mathcal{F}, Prob)$ to represent the IS and OOS performance of the N strategies, respectively. For a given sample $c \in \mathcal{T}$, that is a concrete pair of IS and OOS samples, we will use $\mathbf{R}^c$ and $\overline{\mathbf{R}}^c$ to signify the performances of the N strategies on the IS and OOS pair given by c. For most applications $\mathcal{T}$ will be finite and one can choose to use the power set $\mathcal{T}$ as $\mathcal{F}$. Moreover, often it makes sense in this case to assume that the $Prob$ is uniform on elements in $\mathcal{T}$. However, we do not make specific assumptions at this stage of general discussion so as to allow flexibility in particular applications.

The key observation here is to compare the ranking of the selected strategies IS and OOS. Therefore we consider the ranking space Ω consists of the $N!$ permutations of $(1, 2, \ldots, N)$ indicating the ranking of the N strategies. Then we use random vectors $r, \bar{r}$ to represent the ranking of the components of $\mathbf{R}, \overline{\mathbf{R}}$, respectively. For example, if $N = 3$ and the performance measure is the Sharpe ratio, for a particular sample $c \in \mathcal{T}$, $\mathbf{R}^c = (0.5, 1.1, 0.7)$ and $\overline{\mathbf{R}}^c = (0.6, 0.7, 1.3)$, then we have $r^c = (1, 3, 2)$ and $\bar{r}^c = (1, 2, 3)$. Thus, both r and $\bar{r}$ are random vectors mapping $(\mathcal{T}, \mathcal{F}, Prob)$ to Ω.

Now, we define backtest overfitting, in the context of investment strategy selection alluded to above. We will need to use the following subset of Ω: $\Omega_n^* = \{f \in \Omega \mid f_n = N\}$.

Definition 2.1. (Backtest Overfitting) *We say that the backtest strategy selection process overfits if a strategy with optimal performance IS has an expected ranking below the median OOS. By the Bayesian formula and using the notation above that is*

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

Definition 2.2. (Probability of Backtest Overfitting) *A strategy with optimal performance IS is not necessarily optimal OOS. Moreover, there is a non-null probability that this strategy with optimal performance IS ranks below the median OOS. This is what we define as the* probability of backtest overfit (PBO)*. More precisely,*

$$PBO = \sum_{n=1}^{N} Prob[\overline{r_n} < N/2 \mid r \in \Omega_n^*] Prob[r \in \Omega_n^*]. \qquad (2.2)$$

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 rank 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 used to select the optimal strategy 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.

2.2 The CSCV procedure

The framework Subsection 2.1 is flexible in dealing with the probability of backtest overfitting and other statistical characterizations related to the issue of overfitting. However, in order to quantify say the PBO for concrete applications we need a method to estimate the probability that was abstractly defined in the previous section. Estimating the probability in a particular application relies on schemes for selecting samples of IS and OOS pairs. This section is devoted to establishing such a procedure, which we name *combinatorially symmetric cross-validation*, abbreviated as (CSCV) for convenience of reference.

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 might try alternative sample lengths on which the moving averages are computed, entry thresholds, exit thresholds, stop losses, holding periods, sampling frequencies, and so on. 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.

Algorithm 2.3 (CSCV)**.** We proceed as follows.

First, we form a matrix $\mathbf{M}$ by collecting the performance series from the N trials. In particular, each column $n = 1, \ldots, N$ represents a vector of profits and losses over $t = 1, \ldots, T$ observations associated with a particular model configuration tried by the researcher. $\mathbf{M}$ is therefore a real-valued matrix of order $(T \times N)$. The only conditions we impose are that:

i) $\mathbf{M}$ 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 could be maintained on various slices of the reported performance. If different model configurations trade with different frequencies, observations should be aggregated to match a common index $t = 1, \ldots, T$.

Second, we partition $\mathbf{M}$ across rows, into an even number S of disjoint submatrices of equal dimensions. Each of these submatrices $\mathbf{M}_s$, with $s = 1, \ldots, S$, is of order $(T/S \times N)$.

Third, we form all combinations C_S of $\mathbf{M}_s$, taken in groups of size $S/2$. This gives a total number of combinations

$$\binom{S}{S/2} = \binom{S-1}{S/2-1} \frac{S}{S/2} = \ldots = \prod_{i=0}^{S/2-1} \frac{S-i}{S/2-i} \qquad (2.3)$$

For instance, if $S = 16$, we will form $12,780$ combinations. Each combination $c \in C_S$ is composed of $S/2$ submatrices $\mathbf{M}_s$.

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

a) Form the *training set* J, by joining the $S/2$ submatrices $\mathbf{M}_s$ that constitute c in their original order. J is a matrix of order $(T/S)(S/2) \times N) = T/2 \times N$.

b) Form the *testing set* $\overline{J}$, as the complement of J in M. In other words, $\overline{J}$ is the $T/2 \times N$ matrix formed by all rows of M that are not part of J also in their original order. (The order in forming J and $\overline{J}$ does not matter for some performance measures such as the Sharpe ratio but does matter for others e.g. return maximum drawdown ratio).

c) Form a vector $\mathbf{R}^c$ of performance statistics of order N, where the nth component R^c_n of $\mathbf{R}^c$ reports the performance associated with the nth column of J (the testing set). As before rank of the components of $\mathbf{R}^c$ is denoted by r^c the IS ranking of the N strategies.

d) Repeat c) with J replaced by $\overline{J}$ (the test set) to derive $\overline{\mathbf{R}}^c$ and $\bar{r}^c$ the OOS performance statistics and rank of the N strategies, respectively.

e) Determine the element n^* such that $r^c_{n^*} \in \Omega^*_{n^*}$. In other words, n^* is the best performing strategy IS.

f) Define the relative rank of $\bar{r}^c_{n^*}$ by $\bar{\omega}_c := \bar{r}^c_{n^*}/(N+1) \in (0,1)$. This is the relative rank of the OOS performance associated with the strategy chosen IS. If the strategy optimization procedure is not overfitting, we should observe that $\bar{r}^c_{n^*}$ systematically outperforms OOS, just as $r^c_{n^*}$ outperformed IS.

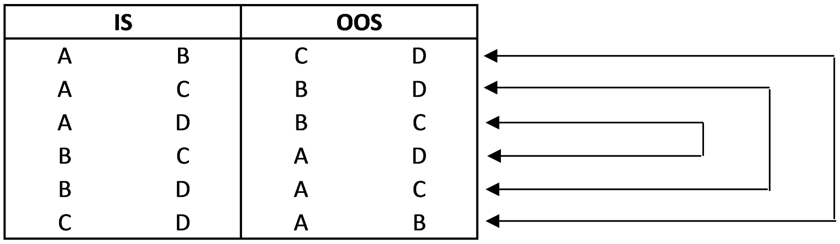
g) We define the *logit* $\lambda_c = \ln \frac{\bar{\omega}_c}{(1-\bar{\omega}_c)}$. High logit values imply a consistency between IS and OOS performances, which indicates a low level of backtest overfitting.

Fifth, we compute the distribution of ranks OOS by collecting all the λ_c, for $c \in C_S$. Define the *relative frequency* at which λ occurred across all C_S by

$$f(\lambda) = \sum_{c \in C_S} \frac{\chi_{\{\lambda\}}(\lambda_c)}{\#(C_S)}, \tag{2.4}$$

where χ is the characterization function and $\#(C_S)$ signifies the number of elements in C_S. Then $\int_{-\infty}^{\infty} f(\lambda) d\lambda = 1$. This concludes the procedure.

Figure 1 schematically represents how the combinations in C_S are used to produce training and testing sets, where $S = 4$. It 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



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

Figure 1: Generating the C_S symmetric combination.

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 vice-versa (as is possible because we split the data in two equal parts).

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 analysis:

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*: This determines to what extent greater performance IS leads to lower performance OOS, an occurrence associated with the memory effects discussed in Bailey et al. [1].

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 backtest overfitting (PBO)

The PBO defined in Section 2.1 may now be estimated using the CSCV method with $\phi = \int_{-\infty}^{0} f(\lambda)d\lambda$. This represents the rate at which optimal IS strategies underperform the median of the OOS trials. The analogue of $\bar{r}$ 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 $\lambda_c > 0$. For $\phi \approx 0$, a low proportion of the optimal IS strategy outperformed the median of trials in most of the testing sets indicating no significant overfitting. On the flip side, $\phi \approx 1$ indicates high likelihood of overfitting. We consider at least three uses for PBO: i) In general the value of ϕ provides us a quantitative sense about the likelihood of overfitting. In accordance with standard applications of the Neyman-Pearson framework, a customary approach would be to reject models for which PBO is estimated to be greater than 0.05. ii) PBO could be used as a prior probability in Bayesian applications, where for instance the goal may be to derive the posterior probability of a model's forecast. iii) We could compute the PBO on a large number of investment strategies, and use those PBO estimates to compute a weighted portfolio, where the weights are given by $(1 - PBO)$, $1/PBO$ or some other scheme.

3.2 Performance degradation and probability of loss

Section 2.2 introduced the procedure to compute, among other results, the pair $(R_{n^*}, \overline{R_{n^*}})$ for each combination $c \in C_S$. Note that while we know that R_{n^*} is the maximum among the components of $\mathbf{R}$, $\overline{R_{n^*}}$ is not necessarily the maximum among the components of $\overline{\mathbf{R}}$. Because we are trying every combination of $\mathbf{M}_s$ taken in groups of size $S/2$, there is no reason to expect the distribution of $\mathbf{R}$ to dominate over $\overline{\mathbf{R}}$. The implication is that, generally, $\overline{R_{n^*}} < \max\{\overline{\mathbf{R}}\} \approx \max\{\mathbf{R}\} = R_{n^*}$. For a regression $\overline{R_{n^*}}^c = \alpha + \beta R_{n^*}^c + \varepsilon^c$, the β will be negative in most practical cases, due to compensation effects described in Bailey et al. [1]. An intuitive explanation for this negative slope is that overfit backtests minimize future performance: The model is so fit to past noise, that it is often rendered unfit for future signal. And the more overfit a backtest is, the more memory is accumulated against its future performance.

It is interesting to plot the pairs $(R_{n^*}, \overline{R_{n^*}})$ to visualize how strong is such performance degradation, andto obtain a more realistic range of attainable performance OOS (see Figure 8). A particularly useful statistic is the proportion of combinations with negative performance, $Prob[\overline{R_{n^*}}^c < 0]$. Note that, even if $\phi \approx 0$, $Prob[\overline{R_{n^*}}^c < 0]$ could be high, in which case the strategy's performance OOS is probably poor for reasons other than overfitting.

Figure 2 provides a graphical representation of i) Out-Of-Sample Performance Degradation, ii) Out-Of-Sample Probability of Loss, and iii) Proba-

Figures

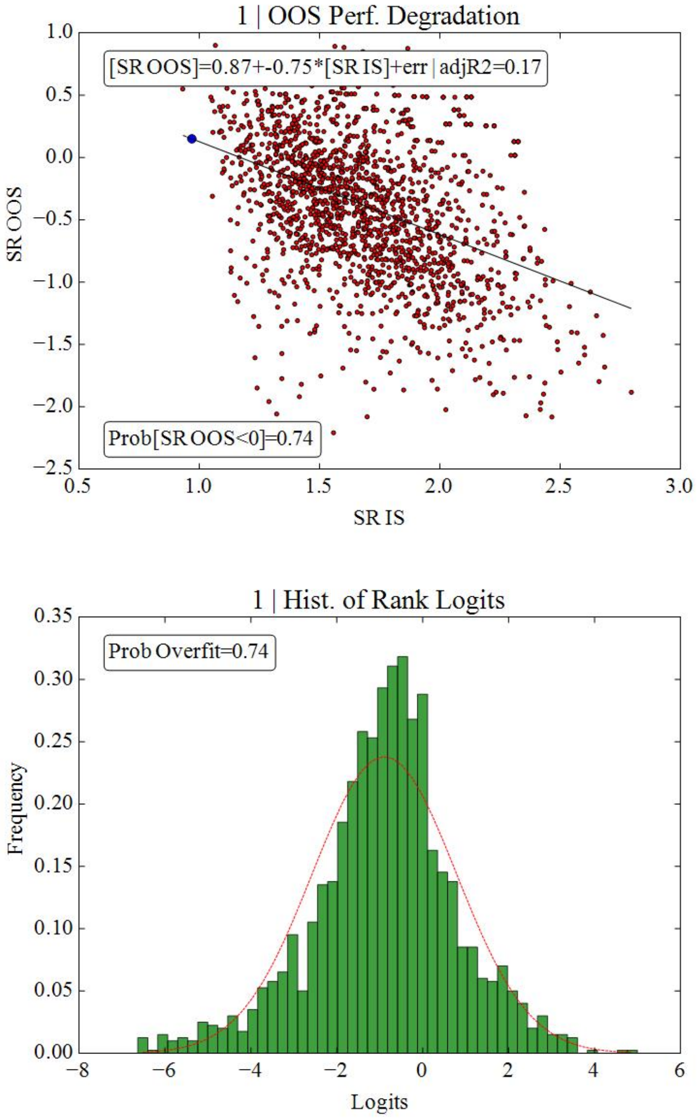


Figure 2: Performance degradation and distribution of logits. Note that even if $\phi \approx 0$, $Prob[\overline{R_{n^*}}^c < 0]$ could be high, in which case the strategy's performance OOS is poor for reasons other than overfitting.

bility of Overfitting (PBO).

The upper plot of Figure 2 shows that pairs of (SR IS, SR OOS) for the optimal model configurations selected for each subset $c \in C_S$, which corresponds to 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, indicating 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, Sharpe ratios 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 again 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.). The method also allows us to compute the proportion of combinations with negative performance, $Prob[\overline{R_{n^*}}^c < 0]$, which corresponds to analysis ii).

The lower plot of Figure 2 shows the distribution of logits for the same strategy, with a PBO of 74%. It displays the distribution of logits, which allows us to compute the probability of backtest overfitting (PBO). This represents the rate at which optimal IS strategies underperform the median of the OOS trials.

Figure 3 plots 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%.

The upper plot of Figure 3 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. The lower plot of Figure 3 shows the distribution of logits, with a PBO of 0.04%. According to this analysis, it is unlikely that this backtest is significantly 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

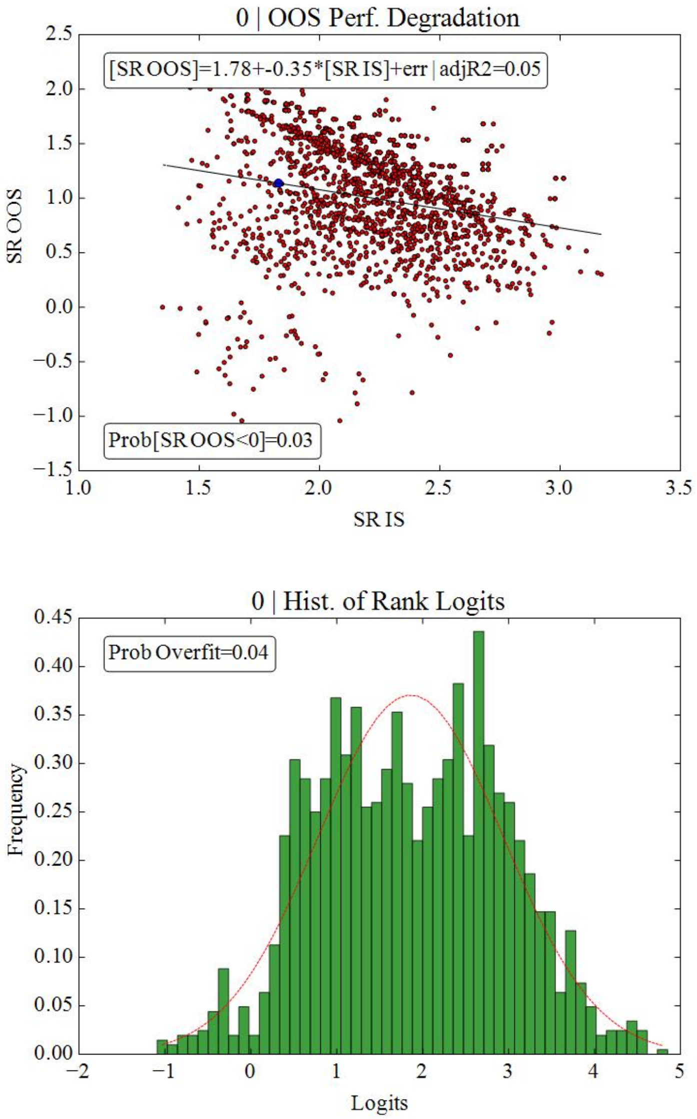


Figure 3: Performance degradation and distribution of logits for a real investment strategy.

over the distribution of all $\overline{\mathbf{R}}$. 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 the concept of stochastic dominance is 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 [11] for an introduction to these matters.

In the context of our framework, first-order stochastic dominance occurs if $Prob[\overline{R_{n^*}} \geq x] \geq Prob[Mean(\overline{\mathbf{R}}) \geq x]$ for all x, and for some x, $Prob[\overline{R_{n^*}} \geq x] > Prob[Mean(\overline{\mathbf{R}}) \geq x]$. It can be verified visually by checking that the cumulative distribution function of $\overline{R_{n^*}}$ is not above the cumulative distribution function of $\mathbf{R}$ 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 $\overline{\mathbf{R}}$, assuming only that her utility function is weakly increasing.

A less demanding criterion is second-order stochastic dominance. This requires that $SD2[x] = \int_{-\infty}^{x} (Prob[Mean(\overline{\mathbf{R}}) \leq x] - Prob[\overline{R_{n^*}} \leq x])dx \geq 0$ 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 $\overline{\mathbf{R}}$, as long as she is risk averse and her utility function is weakly increasing.

Figure 4 complements the analysis presented in Figure 2, with analysis of stochastic dominance. Stochastic dominance allows us to rank gambles or lotteries without having to make strong assumptions regarding an individual's utility function.

Figure 4 also provides an example of the cumulative distribution function of $\overline{R_{n^*}}$ across all $c \in C_S$ (red line) and $\overline{\mathbf{R}}$ (blue line), as well as the second order stochastic dominance ($SD2[x]$, green line) for every OOS SR. In this example, the distribution of OOS SR of optimized (IS) model configurations does not dominate (to first order) the distribution of OOS SR of overall model configurations.

This can be seen 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 even according to this less demanding criterion. It has been computed on the same backtest used for Figure 2. Consistent with that result, the overall distribution of OOS performance dominates the

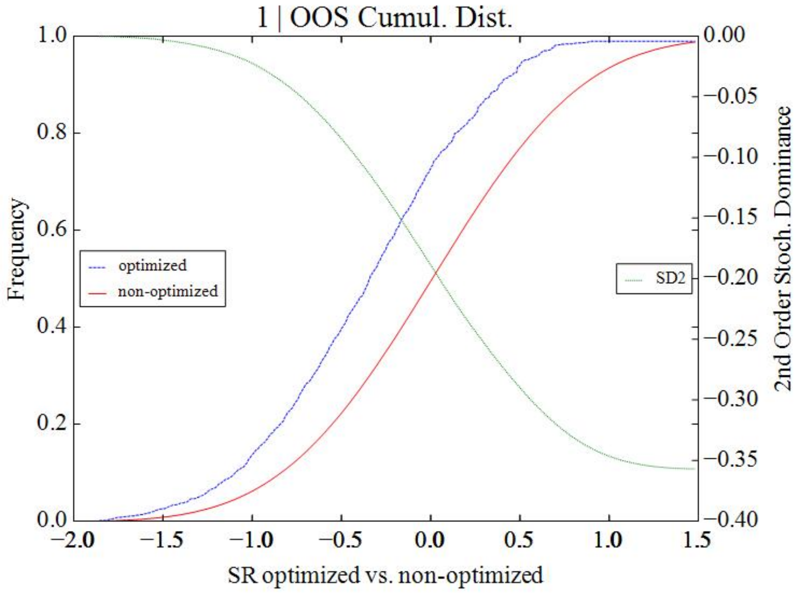


Figure 4: Stochastic dominance (example 1).

OOS performance of the optimal strategy selection procedure, a clear sign of overfitting.

Figure 5 provides a counter-example, based on the same real investment strategy used in Figure 3. It indicates that the strategy selection procedure used in this backtest actually added value, since 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 CSCV sampling method

Our testing method utilises multiple developments 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).

Now, K-FCV randomly divides the sample of size T into k subsamples

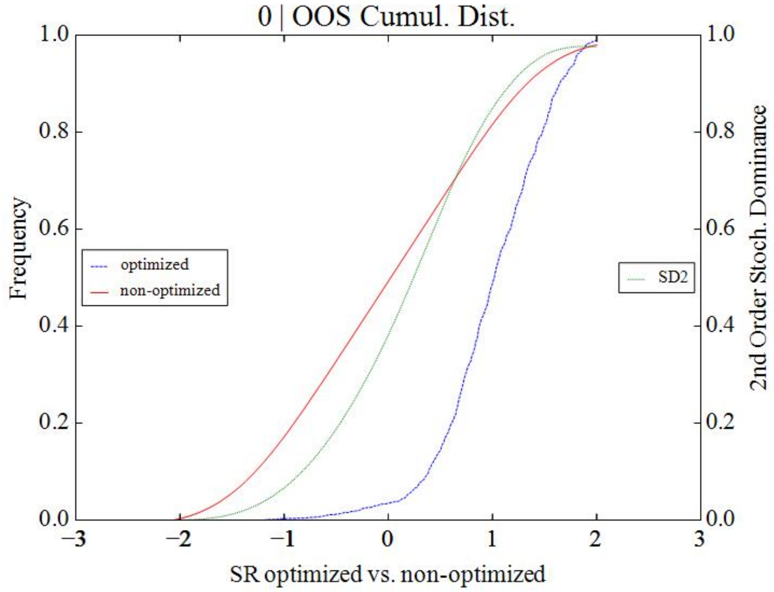


Figure 5: Stochastic dominance (example 2).

of size T/k. Then it sequentially tests on each of the k samples the model trained on the $T - T/k$ sample. Although a very valid approach in many situations, we believe that our procedure is more satisfactory 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 T/k. This means that k must be sufficiently small, so that the Sharpe ratio estimate is reliable (see Bailey and López de Prado [2] 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 unreliable. Also, LOOCV is a K-FCV where $k = T$. We are not aware of any reliable performance metric computed on a single OOS observation.

The *combinatorially symmetric cross-validation* (CSCV) method we have proposed in Section 2.2 differs from both K-FCV and LOOCV. The key idea is to generate $\binom{S}{S/2}$ testing sets of size $T/2$ by recombining the S slices of the overall sample of size T. 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 was 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 season-dependent features 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. As with 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 uniformly (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 M_s that will be generated, each of order $(T/S \times N)$, and also the number of logit values that will be computed, $\binom{S}{S/2}$. Indeed, 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. By 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. The accuracy of the procedure relies on computing a large number of logits, where that number is derived in Equation (2.3). Because $f(\lambda)$ is estimated as a proportion of the number of logits, S needs to be large enough to generate sufficient logits. For a proportion $\hat{p}$ estimated on a sample of size N, the standard deviation of its expected value can be computed as $\sigma[\hat{p}] = \sqrt{\frac{p(1-p)}{N}}$ (see Gelman and Hill [10]). In other words, the standard deviation is highest for $p = \frac{1}{2}$, with $\sigma[\hat{p}] = \sqrt{\frac{1}{4N}}$. Fortunately, even a small number S generates a large number of logits. For example, $S = 16$ we will obtain $12,780$ logits (see Equation (2.3)), and $\sigma[f(\lambda)] < 0.0045$, with less than a 0.01 estimation error at 95% confidence level. Also, if M contains 4 years of daily data, $S = 16$ would equate to quarterly partitions, and the serial correlation structure would be preserved. For these two reasons, we believe that $S = 16$ is a reasonable value to use in most cases.

Another key parameter is the number of trials (i.e., the number of columns in 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. [1] 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 take only a very few values, which will translate into a very discrete number of logits, making $f(\lambda)$ too discontinuous, and adding estimation error to the evaluation of ϕ. For example, if the investor is sensitive to values of $\phi < 1/10$, it is clear that the range of values that the logits can adopt must be greater than 10, and so $N >> 10$ is required. Other considerations regarding N will be discussed in the following Section.

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

5 Limitations and misuse

The general framework in Subsection 2.1 can be flexibly used to assess backtest overfitting probability. Quantitative assessment, however, also relies on methods for estimating the probability measure. In this paper, we focus on one of such implementations: the CSCV method. 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. However, any particular implementation has its limitations and the CSCV method is no exception. Below is a discussion of some of the limitations of this method from the perspective of design and application.

5.1 Limitation in design

First, a key feature of the CSCV implementation is symmetry. In dividing the total sample of the testing results into IS and OOS both the size and method of division in CSCV are symmetric. The advantage of such an symmetric division has been elaborated above. However, the complexity of investment strategies and performance measures makes it unlikely that any particular method will be a one size fits all solution. For some backtests other methods, for example K-FCV, may well be better suited.

Moreover, symmetrically dividing the sample performance in to S symmetrically layered sub-samples also may not suitable for certain strategies. For example, if the performance measure as a time series has a strong autocorrelation, then such a division may obscure the characterization especially when S is large.

Finally, the CSCV estimate of the probability measure assumes all the sample statistics carries the same weight. Without knowing any prior information on the distribution of the backtest performance measure this is, of course, a natural and reasonable choice. If, however, one does have knowledge regarding the distribution of the backtest performance measure, then model-specific methods of dividing the sample performance measure and assigning different weights to different strips of the subdivision are likely to be more accurate. For instance, 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.

5.2 Limitation in application

First, the researcher must provide full information regarding the actual trials conducted, to avoid the file drawer problem (the test is only as good as the completeness of the underlying information), and should test as many strategy configurations as is reasonable and feasible. Hiding trials will lead to an underestimation of the overfit, because each logit will be evaluated under a biased relative rank $\overline{\omega_c}$. This would be 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. A case in point is guided searches, where an optimization algorithm uses information from prior iterations to decide what direction should be followed next. In this case, the columns of matrix M should be the final outcome of each guided search (i.e., after it has converged to a solution), and not the intermediate steps.[2] 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 bad assumptions, such as incorrect 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 a 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.

Fourth, although a high PBO indicates overfitting in the group of N tested strategies, skillful strategies can still exists in these N strategies. For example, it is entirely possible that all the N strategies have high but similar Sharpe ratios. Since none of the strategies is clearly better than the rest, PBO will be high. Here overfitting is among many ‘skillful’ strategies.

Fifth, we must warn the reader against applying CSCV to guide the

[2] We thank David Aronson and Timothy Masters (Baruch College) for asking for this clarification.

search for an optimal strategy. That would constitute a gross misuse of our method. As Strathern [31] eloquently put it, "when a measure becomes a target, it ceases to be a good measure." Any counter-overfitting technique used to select an optimal strategy will result in overfitting. For example, CSCV can be employed to evaluate the quality of a strategy selection process, but PBO should not be the objective function on which such selection relies.

6 A practical application

Bailey et al. [1] 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 $[1, \ldots, 22]$ for Entry_day, the interval $[1, \ldots, 20]$ for Holding_period, the interval $[0, \ldots, 10]$ 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.

First, as discussed in the above cited paper, a time series of $1,000$ 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 of 1.27. Given the elevated Sharpe ratio, we may 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 [2] for details). Figure 6 gives a graphical illustration of this example.

We have estimated the PBO using our CSCV procedure, and obtained the results illustrated below. Figure 7 shows that approx. 53% of the SR OOS are negative, despite all SR IS being positive and ranging between 1 and 2.2. Figure 8 plots the distribution of logits, which implies that, despite the elevated SR IS, the PBO is as high as 55%. Consequently,

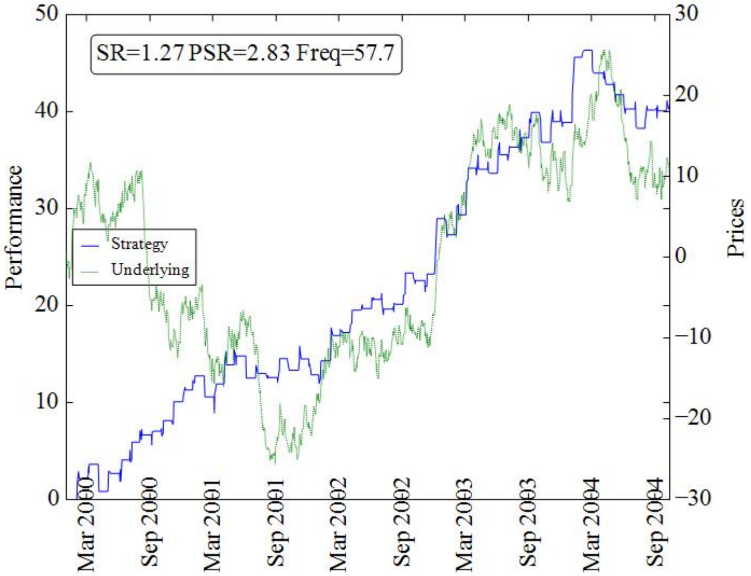


Figure 6: Backtested performance of a seasonal strategy (example 1).

Figure 9 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 diagnosing that the backtest was overfit.

Second, we generated a time series of 1,000 daily prices (about 4 years), following a random walk. But unlike 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 simulates a monthly seasonal effect, which the strategy selection procedure should discover. Figure 10 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 at 1.54 is similar to the previous (overfit) case (1.54 vs. 1.3).

The next three graphs report the results of the CSCV analysis, which confirm the validity of this backtest in the sense that performance inflation from overfitting is minimal. Figure 11 shows only 13% of the OOS SR to be negative. Because there is a real monthly effect in the data, the PBO for

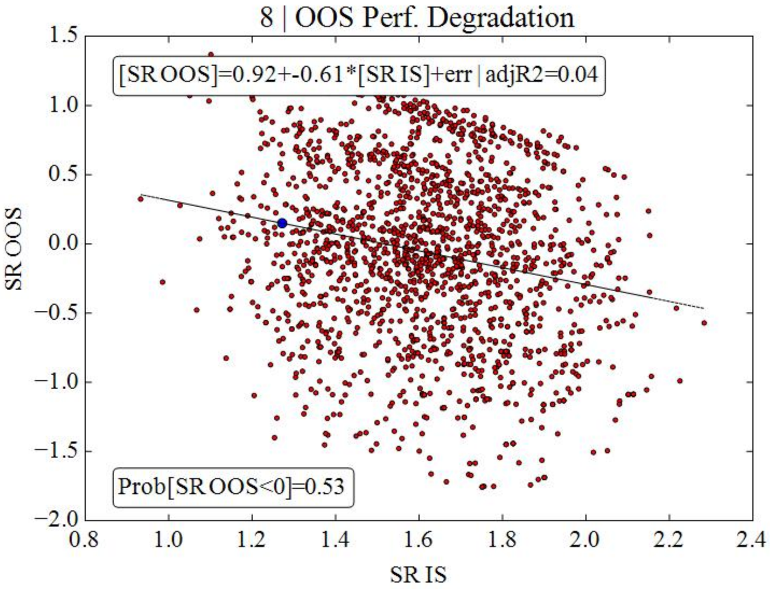


Figure 7: CSCV analysis of the backtest of a seasonal strategy (example 1): Performance degradation.

this second case should be substantially lower than the PBO of the first case. Figure 12 shows a distribution of logits with a PBO of only 13%. Figure 13 evidences that the distribution of OOS SR from IS optimal combinations clearly dominates the overall distribution of OOS SR. The CSCV analysis has this time correctly recognized the validity of this backtest, in the sense that performance inflation from overfitting is small.

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 [1, 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, and the like. 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 results.

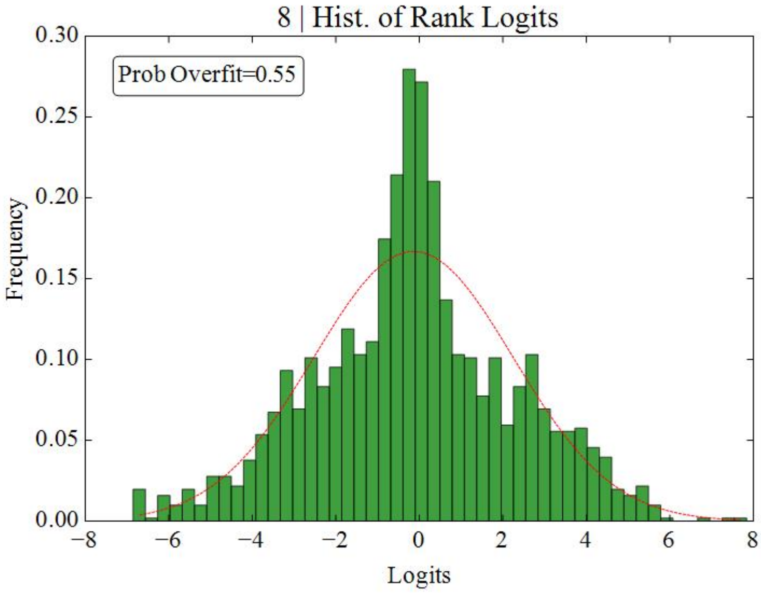


Figure 8: CSCV analysis of the backtest of a seasonal strategy (example 1): logit distrubution.

7 Conclusions

In [2] Bailey and López de Prado developed methodologies to evaluate the probability that a Sharpe ratio is inflated (PSR), and to determine 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 investment performance and backtest track records. This paper has extended this approach to present formulas and approximation techniques for finding the probability of backtest overfitting.

To that end, we have proposed a general framework for modeling the IS and OOS performance using probability. We define the probability of backtested overfitting (PBO) as the probability that an optimal strategy IS underperforms the mean OOS. To facilitate the evaluation of PBO for particular applications, we have proposed a combinatorially symmetric cross-validation (CSCV) implementation framework for estimating this probability. This estimate is generic, symmetric, model-free and non-parametric. We have assessed the accuracy of CSCV as an approximation of PBO in

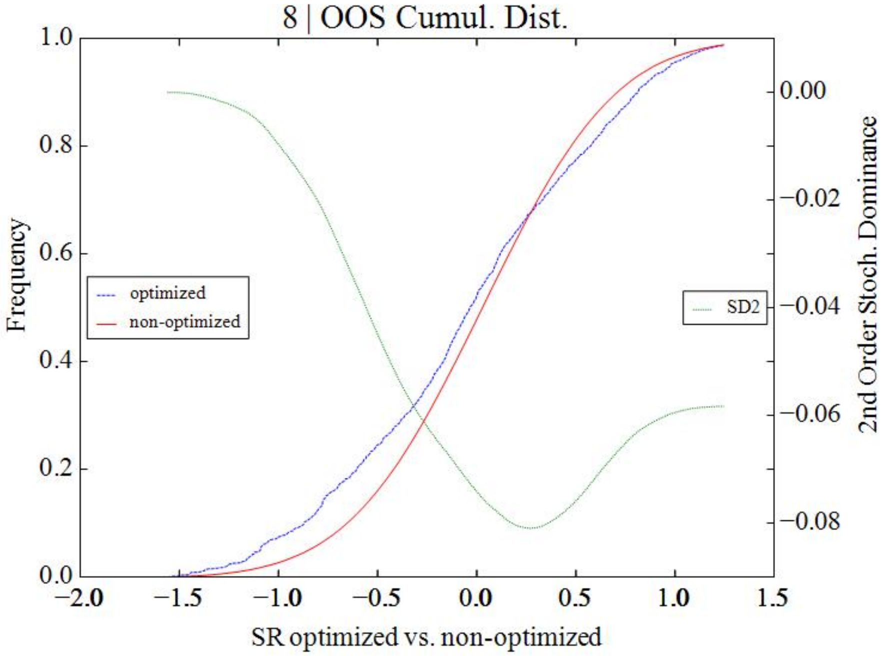


Figure 9: CSCV analysis of the backtest of a seasonal strategy (example 1): Absent of dominance.

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 directly computed PBO by deriving the Extreme Value distributions that model the performance of IS optimal strategies. These results indicate that CSCV provides reasonable estimates of PBO, with relatively small errors.

Besides estimating PBO, our general framework and its CSCV implementation scheme can also be used to deal with other issues related to overfitting, such as performance degeneration, probability of loss and possible stochastic dominance of a strategy. On the other hand, the CSCV implementation also has some limitations. This suggests that other implementation frameworks may well be more suitable, particularly for problems with structure information.

Nevertheless, we believe that CSCV provides both a new and powerful tool in the arsenal of an investment and financial researcher, and that it also

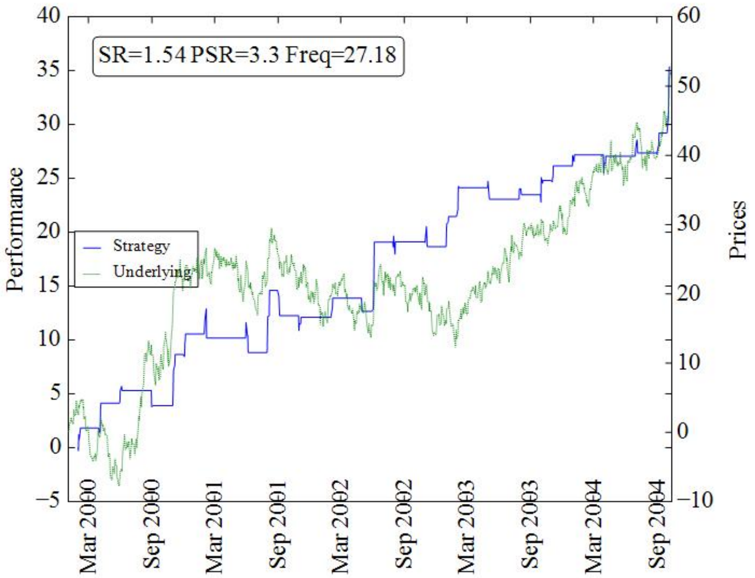


Figure 10: Backtested performance of a seasonal strategy (example 2).

constitutes a nice illustration of our general framework for quantitatively studying issues related to backtest overfitting. We certainly hope that this study will raise greater awareness concerning the futility of computing and reporting backtest results, without first controlling for PBO and MinBTL.

References

[1] Bailey, D., J. Borwein, M. López de Prado and J. Zhu, "Pseudo-mathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance," *Notices of the AMS*, **61** May (2014), 458–471. Online at `http://www.ams.org/notices/201405/rnoti-p458.pdf`.

[2] Bailey, D. and M. López de Prado, "The Sharpe Ratio Efficient Frontier," *Journal of Risk*, 15(2012), 3–44. Available at `http://ssrn.com/abstract=1821643`.

[3] Bailey, D. and M. López de Prado, "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality", *Journal of Portfolio Management*, 40 (5) (2014), 94-107.

[4] Calkin, N. and M. López de Prado, "Stochastic Flow Diagrams", *Algorithmic Finance*, 3(1-2) (2014) Available at `http://ssrn.com/abstract=2379314`.

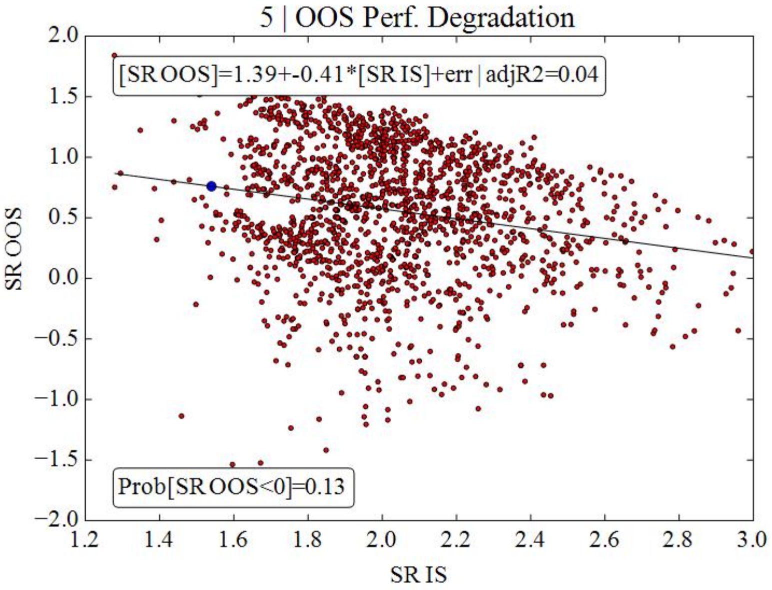


Figure 11: CSCV analysis of the backtest of a seasonal strategy (example 2): monthly effect.

[5] Calkin, N. and M. López de Prado, "The Topology of Macro Financial Flows: An Application of Stochastic Flow Diagrams", *Algorithmic Finance*, 3(1-2) (2014). Available at `http://ssrn.com/abstract=2379319`.

[6] Carr, P. and M. López de Prado, "Determining Optimal Trading Rules without Backtesting", (2014) Available at `http://arxiv.org/abs/1408.1159`.

[7] Doyle, J. and C. Chen, "The wandering weekday effect in major stock markets," *Journal of Banking and Finance*, 33 (2009), 1388–1399.

[8] Embrechts, P., C. Klueppelberg and T. Mikosch, *Modelling Extremal Events*, Springer Verlag, New York, 2003.

[9] Feynman, R., *The Character of Physical Law*, 1964, The MIT Press.

[10] Gelman, A. and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2006, Cambridge University Press, First Edition.

[11] Hadar, J. and W. Russell, "Rules for Ordering Uncertain Prospects," *American Economic Review*, 59 (1969), 25–34.

[12] Harris, L., *Trading snf Exchanges: Market Microstructure for Practitioners*, Oxford University Press, 2003.

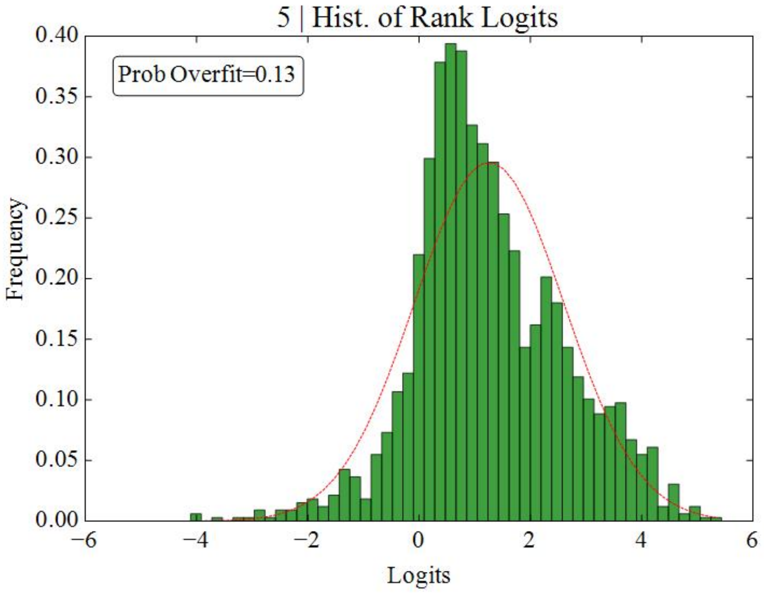


Figure 12: CSCV analysis of the backtest of a seasonal strategy (example 2): logit distribution.

[13] Harvey, C. and Y. Liu, "Backtesting", SSRN, working paper, 2013. Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2345489`.

[14] Harvey, C., Y. Liu and H. Zhu, "...and the Cross-Section of Expected Returns," SSRN, 2013. Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2249314`.

[15] Hawkins, D., "The problem of overfitting," *Journal of Chemical Information and Computer Science*, 44 (2004), 10–12.

[16] Hirsch, Y., *Don't Sell Stocks on Monday*, Penguin Books, 1st Edition, 1987.

[17] Ioannidis, J.P.A., "Why most published research findings are false." PloS Medicine, Vol. 2, No. 8,(2005) 696-701.

[18] Leinweber, D. and K. Sisk,"Event Driven Trading and the 'New News'," *Journal of Portfolio Management*, 38(2011), 110–124.

[19] Leontief, W., "Academic Economics", *Science*, 9 Jul 1982, 104–107.

[20] Lo, A., "The Statistics of Sharpe Ratios," *Financial Analysts Journal*, 58 (2002), July/August.

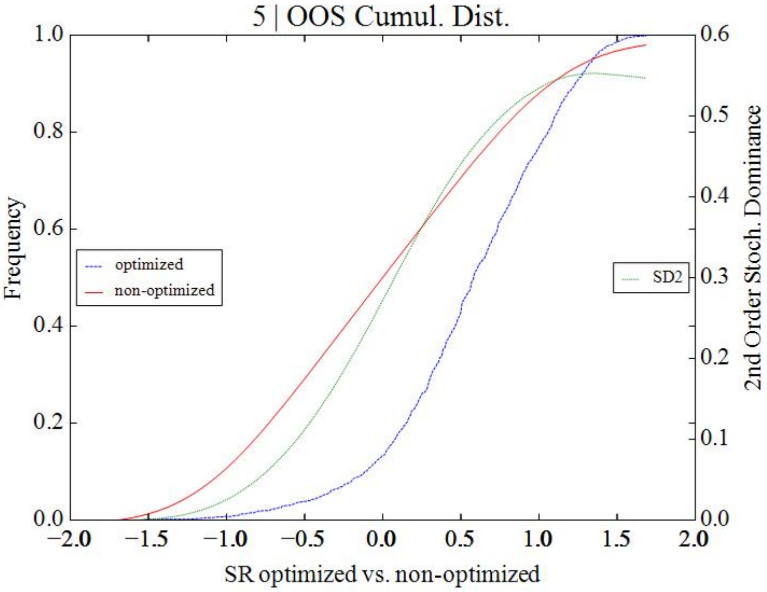


Figure 13: CSCV analysis of the backtest of a seasonal strategy (example 2): dominance.

[21] López de Prado, M. and A. Peijan, “Measuring the Loss Potential of Hedge Fund Strategies,” *Journal of Alternative Investments*, 7 (2004), 7–31. Available at `http://ssrn.com/abstract=641702`.

[22] López de Prado, M. and M. Foreman, ”A Mixture of Gaussians Approach to Mathematical Portfolio Oversight: The EF3M Algorithm”, Quantitative Finance, forthcoming, 2014. Available at `http://ssrn.com/abstract=1931734`.

[23] MacKay, D.J.C. “Information Theory, Inference and Learning Algorithms”, Cambridge University Press, First Edition, 2003.

[24] Mayer, J., K. Khairy and J. Howard, “Drawing an Elephant with Four Complex Parameters,” *American Journal of Physics*, 78 (2010), 648–649.

[25] Miller, R.G., *Simultaneous Statistical Inference*, 2nd Ed. Springer Verlag, New York, 1981. ISBN 0-387-90548-0.

[26] Resnick, S., *Extreme Values, Regular Variation and Point Processes*, Springer, 1987.

[27] Romano, J. and M. Wolf, “Stepwise multiple testing as formalized data snooping”, *Econometrica*, 73 (2005), 1273–1282.

[28] Sala-i-Martin, X., "I just ran two million regressions." *American Economic Review.* 87(2), May (1997).

[29] Schorfheide, F. and K. Wolpin, "On the Use of Holdout Samples for Model Selection," *American Economic Review*, 102 (2012), 477–481.

[30] Stodden, V., Bailey, D., Borwein, J., LeVeque, R, Rider, W. and Stein, W., "Setting the default to reproducible: Reproduciblity in computational and experimental mathematics," February, 2013. Available at `http://www.davidhbailey.com/dhbpapers/icerm-report.pdf`.

[31] Strathern, M., "Improving Ratings: Audit in the British University System," European Review, 5, (1997) pp. 305-308.

[32] The Economist, "Trouble at the lab", Oct. 2013 Available at http://www.economist.com/news/briefing/21588057 -scientists -think-science-self -correcting-alarming-degree-it-not-trouble.

[33] Van Belle, G. and K. Kerr, *Design and Analysis of Experiments in the Health Sciences*, John Wiley and Sons, 2012.

[34] Weiss, S. and C. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*, Morgan Kaufman, 1st Edition, 1990.

[35] White, H., "A Reality Check for Data Snooping," *Econometrica*, 68 (2000), 1097–1126.

[36] Wittgenstein, L.: *Philosophical Investigations*, 1953. Blackwell Publishing. Section 201.