PSEUDO-MATHEMATICS AND FINANCIAL CHARLATANISM: THE EFFECTS OF BACKTEST OVERFITTING ON OUT-OF-SAMPLE PERFORMANCE

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

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

April 1, 2014

First version: September 2013

---

[*]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 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/`

Abstract

Recent computational advances allow investment managers to methodically search through thousands or even millions of potential options for a profitable investment strategy. In many instances, the resulting strategy involves a pseudo-mathematical argument, which is spuriously validated through a simulation of its historical performance (also called backtest).

We prove that high performance is easily achievable after backtesting a relatively small number of alternative strategy configurations, a practice we denote "backtest overfitting." The higher the number of configurations tried, the greater is the probability that the backtest is overfit. Because financial analysts rarely report the number of configurations tried for a given backtest, investors cannot evaluate the degree of overfitting in most investment claims and analysis.

The implication is that investors can be easily misled into allocating capital to strategies that appear to be mathematically sound and empirically supported by a backtest. This practice is particularly pernicious, because due to the nature of financial time series, backtest overfitting has a detrimental effect on the strategy's future performance.

Another thing I must point out is that you cannot prove a vague theory wrong. [...] Also, if the process of computing the consequences is indefinite, then with a little skill any experimental result can be made to look like the expected consequences.

Richard Feynman [1964]

1 Introduction

A *backtest* is a historical simulation of an algorithmic investment strategy. Among other things, it computes the series of profits and losses that such strategy would have generated, should that algorithm had been run over that time period. Popular performance statistics, such as the *Sharpe ratio* or the *Information ratio*, are used to quantify the backtested strategy's return on risk. Investors typically study those backtest statistics, and then allocate capital to the best performing scheme.

With regards to the measured performance of a backtested strategy, we have to distinguish between two very different readings: *in-sample* (IS) and *out-of-sample* (OOS). The IS performance is the one simulated over the sample used in the design of the strategy (also known as "learning period" or "training set" in the machine learning literature). The OOS performance is simulated over a sample not used in the design of the strategy (a.k.a. "testing set"). A backtest is *realistic* when the IS performance is consistent with the OOS performance.

When an investor receives a promising backtest from a researcher or portfolio manager, one of her key problems is to assess how realistic that simulation is. This is because, given any financial series, it is relatively simple to *overfit* an investment strategy so that it performs well IS.

Overfitting is a concept borrowed from machine learning, and denotes the situation when a model targets particular observations rather than a general structure. For example, a researcher could design a trading system based on some parameters that target the removal of specific recommendations that she knows led to losses IS (a practice known as "data snooping"). After a few iterations, the researcher will come up with "optimal parameters," which profit from features that are present in that particular sample but may well be rare in the population.

Recent computational advances allow investment managers to methodically search through thousands or even millions of potential options for a profitable investment strategy. In many instances, that search involves a pseudo-mathematical argument, which is spuriously validated through a backtest. For example, consider a time series of daily prices for a stock X.

3

For every day in the sample, we can compute one average price of that stock using the previous m observations $\bar{x}_m$, and another average price using the previous n observations $\bar{x}_n$, where $m < n$. A popular investment strategy called "crossing moving averages" consists in owning X whenever $\bar{x}_m > \bar{x}_n$. Indeed, since the sample size determines a limited number of parameter combinations that m and n can adopt, it is relatively easy to determine the pair (m, n) that maximizes the backtest's performance. There are hundreds of such popular strategies, marketed to unsuspecting lay investors as mathematically sound and empirically tested.

In the context of Econometric models several procedures have been proposed to determine overfit in White [27], Romano et al. [23], Harvey et al. [9]. These methods propose to adjust the p-values of estimated regression coefficients to account for the multiplicity of trials. These approaches are valuable for dealing with trading rules based on an econometric specification.

The machine learning literature has devoted significant effort to study the problem of overfitting. Their proposed methods typically are not applicable to investment problems, for multiple reasons. First, these methods often require explicit point forecasts and confidence bands over a defined event horizon, in order to evaluate the explanatory power or quality of the prediction (e.g., "E-mini S&P500 is forecasted to be around 1,600 with a one-standard deviation of 5 index points at Friday's close"). Very few investment strategies yield such explicit forecasts; instead, they provide qualitative recommendations (e.g., "buy" or "strong buy") over an undefined period until another such forecast is generated, with random frequency. For instance, trading systems, like the crossing of moving averages explained earlier, generate buy and sell recommendations with little or no indication as to forecasted values, confidence on a particular recommendation or expected holding period.

Second, even if a particular investment strategy relies on such a forecasting equation, other components of the investment strategy may have been overfitted, including entry thresholds, risk sizing, profit-taking, stop-loss, cost of capital, and so on. In other words, there are many ways to overfit an investment strategy other than simply tuning the forecasting equation. Third, regression overfitting methods are parametric and involve a number of assumptions regarding the underlying data which may not be easily ascertainable. Fourth, some methods do not control for the number of trials attempted.

To illustrate this point, suppose that a researcher is given a finite sample, and told that she needs to come up with a strategy with a SR (Sharpe Ratio, a popular measure of performance in the presence of risk) above

4

2.0, based on a forecasting equation for which the AIC statistic (Akaike Information Criterion, a standard of the regularization method) rejects the null hypothesis of overfitting with a 95% confidence level (i.e. a false positive rate of 5%). After only 20 trials, the researcher is expected to find one specification that passes the AIC criterion. The researcher will quickly be able to present a specification that not only (falsely) passes the AIC test but it also gives a SR above 2.0. The problem is, AIC's assessment did not take into account the hundreds of other trials that the researcher neglected to mention. For these reasons, commonly used regression overfitting methods are poorly equipped to deal with backtest overfitting.

Although there are many academic studies that claim to have identified profitable investment strategies, their reported results are almost always based on IS statistics. Only exceptionally do we find an academic study that applies the "hold-out" method or some other procedure to evaluate performance OOS. Harvey, Liu and Zhu [10] argue that there are hundreds of papers supposedly identifying hundreds of factors with explanatory power over future stocks returns. They echo Ioannidis [13] in concluding that "most claimed research findings are likely false". Factor models are only the tip of the iceberg.[1] The reader is probably familiar with many publications solely discussing IS performance.

This situation is, quite frankly, depressing, particularly because academic researchers are expected to recognize the dangers and practice of overfitting. One common criticism, of course, is the credibility problem of "holding-out" when the researcher had access to the full sample anyway. Leinweber and Sisk [15] present a meritorious exception. They proposed an investment strategy in a conference, and announced that six months later they would publish the results with the pure (yet to be observed) OOS data. They called this approach "model sequestration," which is an extreme variation of "hold-out."

1.1 OUR INTENTIONS

In this paper we shall show that it takes a relatively small number of trials to identify an investment strategy with a spuriously high backtested performance. We also compute the *minimum backtest length* (MinBTL) that an investor should require given the number of trials attempted. Although in our examples we always choose the Sharpe ratio to evaluate performance,

---

[1]We invite the readers to read specific instances of pseudo-mathematical financial advice at this website: `http://www.m-a-f-f-i-a.org/`. Also, Edesses [2007] provides numerous examples.

our methodology can be applied to any other performance measure.

We believe our framework to be helpful to the academic and investment communities by providing a benchmark methodology to assess the reliability of a backtested performance. We would feel sufficiently rewarded in our efforts if at least this paper succeeded in drawing the attention of the mathematical community regarding the widespread proliferation of journal publications, many of them claiming profitable investment strategies on the sole basis of IS performance. This is perhaps understandable in business circles, but a higher standard is and should be expected from an academic forum.

We would also like to raise the question of whether mathematical scientists should continue to tolerate the proliferation of investment products that are misleadingly marketed as mathematically sound. In the recent words of Sir Andrew Wiles,

> One has to be aware now that mathematics can be misused and that we have to protect its good name. [29]

We encourage the reader to search the Internet for terms such as "stochastic oscillators," "Fibonacci ratios," "cycles," "Elliot wave," "Golden ratio," "parabolic SAR," "pivot point," "momentum," and others in the context of finance. Although such terms clearly evoke precise mathematical concepts, in fact, in almost all cases, their usage is scientifically unsound.

Historically scientists have led the way in exposing those who utilize pseudoscience to extract a commercial benefit. As early as the 18th century, physicists exposed the nonsense of astrologers. Yet mathematicians in the 21st century have remained disappointingly silent with the regards to those in the investment community who, knowingly or not, misuse mathematical techniques such as probability theory, statistics and stochastic calculus. Our silence is consent, making us accomplices in these abuses.

The rest of our study is organized as follows: Section 2 introduces the problem in a more formal way. Section 3 defines the concept of Minimum Backtest Length (MinBTL). Section 4 argues how model complexity leads to backtest overfitting. Section 5 analyzes overfitting in absence of compensation effects. Section 6 studies overfitting in presence of compensation effects. Section 7 exposes how backtest overfitting can be used to commit fraud. Section 8 presents a typical example of backtest overfitting. Section 9 lists our conclusions. The mathematical appendices supply proofs of the propositions presented throughout the paper.

2 BACKTEST OVERFITTING

The design of an investment strategy usually begins with a *prior* or *belief* that a certain pattern may help forecast the future value of a financial variable. For example, if a researcher recognizes a lead-lag effect between various tenor bonds in a yield curve, she could design a strategy that bets on a reversion towards equilibrium values. This model might take the form of a cointegration equation, a vector-error correction model or a system of stochastic differential equations, just to name a few. The number of possible model configurations (or trials) is enormous, and naturally the researcher would like to select the one that maximizes the performance of the strategy. Practitioners often rely on *historical simulations* (also called backtests) to discover the optimal specification of an investment strategy. The researcher will evaluate, among other variables, what are the optimal sample sizes, signal update frequency, entry and profit taking thresholds, risk sizing, stop losses, maximum holding periods, etc.

The *Sharpe ratio* is a statistic that evaluates an investment manager or strategy's performance on the basis of a sample of past returns. It is defined as the ratio between average excess returns (in excess of the rate of return paid by a risk-free asset, such as a Government Note) and the standard deviation of the same returns. Intuitively, this can be interpreted as a "return on risk" (or as William Sharpe put it, "return on variability"). But the standard deviation of excess returns may be a misleading measure of variability when returns follow asymmetric or fat-tailed distributions, or returns are not independent or identically distributed. Suppose that a strategy's excess returns (or risk premiums), r_t, are independent and identically distributed (IID) following a Normal law.

$$r_t \sim \mathcal{N}(\mu, \sigma^2) \tag{2.1}$$

where $\mathcal{N}$ represents a Normal distribution with mean μ and variance σ^2. The *annualized* Sharpe ratio (SR) can be computed as

$$SR = \frac{\mu}{\sigma}\sqrt{q} \tag{2.2}$$

where q is the number of returns per year (see Lo [17] for a detailed derivation of this expression). Sharpe ratios are typically expressed in annual terms in order to allow for the comparison of strategies that trade with different frequency. The great majority of financial models are built upon the IID Normal assumption, which may explain why the Sharpe ratio has become the most popular statistic for evaluating an investment's performance.

Since μ, σ are usually unknown, the true value SR cannot be known for certain. Instead, we can estimate the Sharpe ratio as $\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}}\sqrt{q}$, where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and sample standard deviation. The inevitable consequence is that SR calculations are likely to be the subject of substantial estimation errors (see Bailey and López de Prado [2] for a confidence band, and an extension of the concept of Sharpe Ratio beyond the IID Normal assumption).

From Lo [17], we know that the distribution of the estimated annualized Sharpe ratio $\widehat{SR}$ converges asymptotically (as $y \to \infty$) to

$$\widehat{SR} \xrightarrow{a} \mathcal{N} \left[SR, \frac{1 + \frac{SR^2}{2q}}{y} \right] \tag{2.3}$$

where y is the number of years used to estimate $\widehat{SR}$.[2] As y increases without bound, the probability distribution of $\widehat{SR}$ approaches a Normal distribution with mean SR and variance $\frac{1 + \frac{SR^2}{2q}}{y}$. For a sufficiently large y, Eq. (2.3) provides an approximation of the distribution of $\widehat{SR}$.

Even for a small number N of trials it is relatively easy to find a strategy with a high Sharpe ratio IS, but which also delivers a null Sharpe ratio OOS. To illustrate this point, consider N strategies with $T = yq$ returns distributed according to a Normal law with mean excess returns μ and with standard deviation σ. Suppose that we would like to select the strategy with optimal $\widehat{SR}$ IS, based on one year of observations. A risk we face is choosing a strategy with a high Sharpe ratio IS, but zero Sharpe ratio OOS. So we ask the question, *how high is the expected maximum Sharpe ratio IS among a set of strategy configurations, where the true Sharpe ratio is zero?*

Bailey and López de Prado [2] derived an estimate of the *Minimum Track Record Length (MinTRL)* needed to reject the hypothesis that an estimated Sharpe ratio is below a certain threshold (let's say zero). MinTRL was developed to evaluate a strategy's track record (a single realized path, $N = 1$). The question we are asking now is different, because we are interested in the backtest length needed to avoid selecting a skill-less strategy among N alternative specifications. In other words, in this article we are concerned with overfitting prevention when comparing multiple strategies, not in evaluating the statistical significance of a single Sharpe ratio estimate. Next, we

---

[2]Most performance statistics assume IID Normal returns, and so are normally distributed. In the case of the Sharpe ratio, several authors have proved that its asymptotic distribution follows a Normal law even when the returns are not IID Normal. The same result applies to the Information Ratio. The only requirement is that the returns be ergodic. We refer the interested reader to Bailey and López de Prado [2].

will derive the analogue to MinTRL in the context of overfitting, which we will call *Minimum Backtest Length (MinBTL)*, since it specifically addresses the problem of backtest overfitting.

From Eq. (2.3), if $\mu = 0$ and $y = 1$, then $\widehat{SR} \xrightarrow{a} \mathcal{N}(0,1)$. Note that, because $SR = 0$, increasing q does not reduce the variance of the distribution. The proof of the following proposition is left for the Appendix.

Proposition 2.1. *Given a sample of IID random variables, $x_n \sim Z$, $n = 1, \ldots, N$ where Z is the CDF of the Standard Normal distribution, the expected maximum of that sample, $E[\max_N] = E[\max\{x_n\}]$, can be approximated for a large N as*

$$E[\max_N] \approx (1 - \gamma)Z^{-1}\left[1 - \frac{1}{N}\right] + \gamma Z^{-1}\left[1 - \frac{1}{N}e^{-1}\right] \tag{2.4}$$

where $\gamma \approx 0.5772156649\ldots$ is the Euler-Mascheroni constant, and $N >> 1$.

An upper bound to Eq.(2.4) is $\sqrt{2\ln[N]}$ [3]. Figure 1 plots, for various values of N (x-axis), the expected Sharpe ratio of the optimal strategy IS. For example, if the researcher tries only $N = 10$ alternative configurations of an investment strategy, he or she is expected to find a strategy with a Sharpe ratio IS of 1.57, despite the fact that all strategies are expected to deliver a Sharpe ratio of zero OOS (including the "optimal" one selected IS).

Proposition 2.1 has important implications. As the researcher tries a growing number of strategy configurations, there will be a non-null probability of selecting IS a strategy with null expected performance OOS. Because the hold-out method does not take into account the number of trials attempted before selecting a model, it cannot assess the representativeness of a backtest.

3 MINIMUM BACKTEST LENGTH (MinBTL)

Let us consider now the case that $\mu = 0$ but that $y \neq 1$. Then, we can still apply Proposition 1, by re-scaling the expected maximum by the standard deviation of the annualized Sharpe ratio, $y^{-1/2}$. Thus, the researcher is expected to find an "optimal" strategy with an IS annualized Sharpe ratio of

$$E[\max_N] \approx y^{-1/2}\left((1 - \gamma)Z^{-1}\left[1 - \frac{1}{N}\right] + \gamma Z^{-1}\left[1 - \frac{1}{N}e^{-1}\right]\right) \tag{3.1}$$

---

[3]See Example 3.5.4 of Embrechts et al. [5] for a detailed treatment of the derivation of upper bounds on the maximum of a Normal distribution.
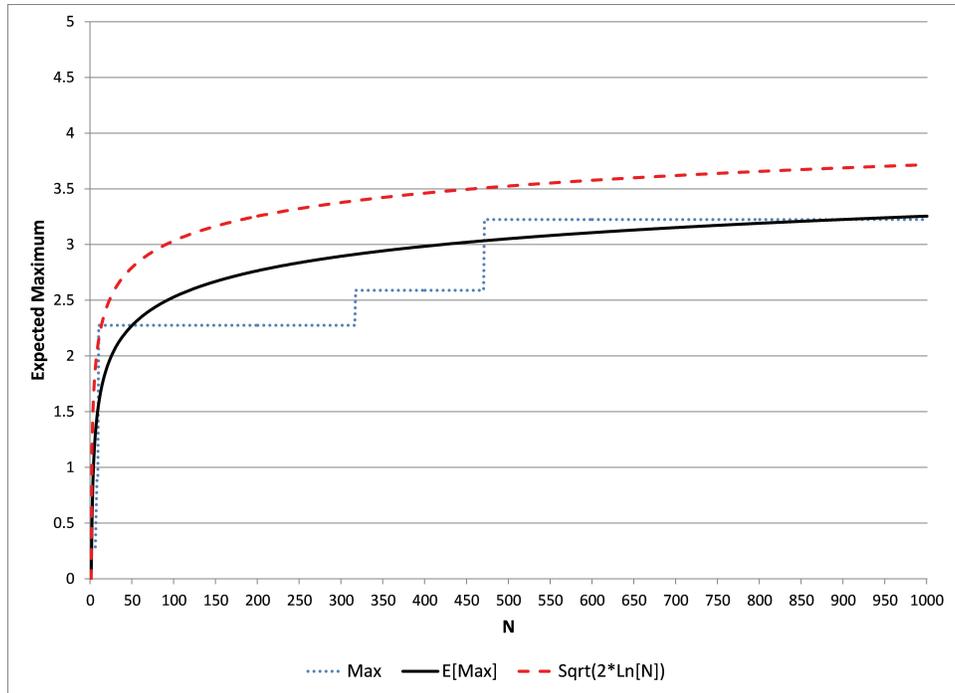
Figure 1: Overfitting a backtest results as the number of trials grows

Figure 1 provides a graphical representation of Proposition 2.1. The blue (dotted) line shows the maximum of a particular set of N independent random numbers, each following a Standard Normal distribution. The black (continuous) line is the expected value of the maximum of that set of N random numbers. The red (dashed) line is an upper bound estimate of that maximum. The implication is that is it relatively easy to wrongly select a strategy on the basis of a maximum Sharpe ratio when displayed IS.

Eq. (3.1) says that the more independent configurations a researcher tries (N), the more likely she is to overfit, and therefore the higher should the acceptance threshold should be for the backtested result to be trusted. This situation can be partially mitigated by increasing the sample size (y). By solving Eq. (3.1) for y, we reach the following statement.

Theorem 3.1. *The Minimum Backtest Length (MinBTL, in years) needed to avoid selecting a strategy with an IS Sharpe ratio of $\overline{E[\max_N]}$ among N independent strategies with an expected OOS Sharpe ratio of zero is*

$$MinBTL \approx \left(\frac{(1-\gamma)Z^{-1}\left[1-\frac{1}{N}\right] + \gamma Z^{-1}\left[1-\frac{1}{N}e^{-1}\right]}{\overline{E[\max_N]}} \right)^2 < \frac{2\ln[N]}{\overline{E[\max_N]}^2} \quad (3.2)$$

Eq. (3.2) tells us that *MinBTL* must grow as the researcher tries more independent model configurations (N), in order to keep constant the expected maximum Sharpe ratio at a given level $\overline{E[\max_N]}$. Figure 2 shows how many years of backtest length (*MinBTL*) are needed so that $E[\max_N]$ is fixed at 1. For instance, if only 5 years of data are available, no more than 45 independent model configurations should be tried, or we are almost guaranteed to produce strategies with an annualized Sharpe ratio IS of 1, but an expected Sharpe ratio OOS of zero. Note that Proposition 2.1 assumed the N trials to be independent, which leads to a quite conservative estimate. If the trials performed were not independent, the number of independent trials N involved can be derived using a dimension-reduction procedure, such as Principal Component Analysis.

We will examine this trade-off between N and T in greater depth later in the paper, without requiring such strong assumption, but *MinBTL* gives us a first glance at how easy is to overfit by merely trying alternative model configurations. As an approximation, the reader may find helpful to remember the upper bound to the minimum backtest length (in years), $MinBTL < \frac{2\ln[N]}{\overline{E[\max_N]}^2}$.

Of course, a backtest may be overfit even if it is computed on a sample greater than MinBTL. From that perspective, MinBTL should be considered a necessary, non-sufficient condition to avoid overfitting. We leave to Bailey et al. [1] the derivation of a more precise measure of backtest overfitting.

4 MODEL COMPLEXITY

How does the previous result relate to model complexity? Consider a one-parameter model that may adopt two possible values (like a switch that
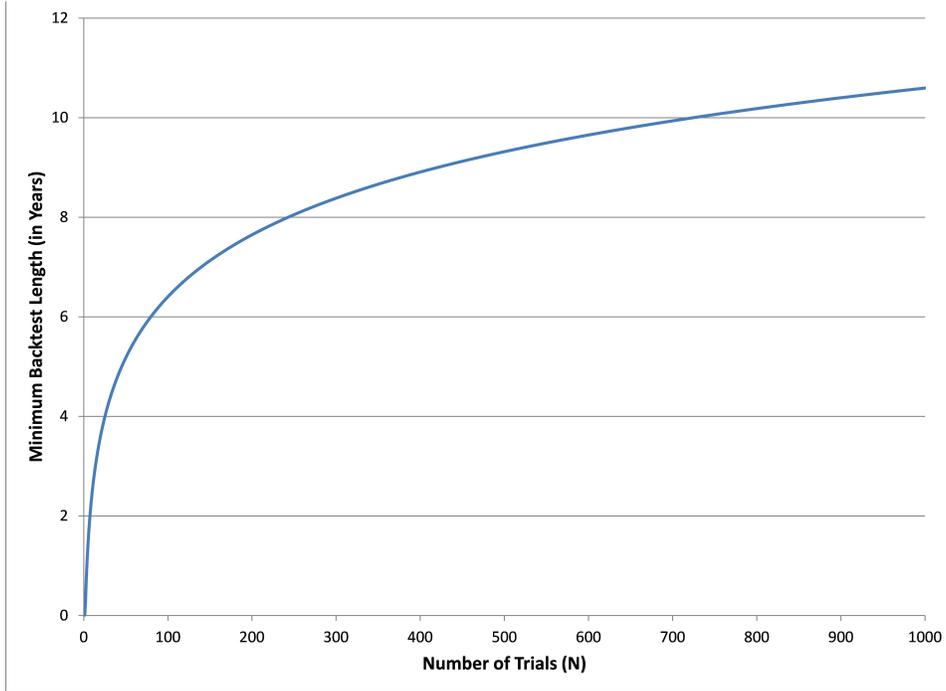
Figure 2: Minimum Backtest Length needed to avoid overfitting, as a function of the number of trials

Figure 2 shows the trade-off between the number of trials (N) and the minimum backtest length (MinBTL) needed to prevent skill-less strategies to be generated with a Sharpe ratio IS of 1. For instance, if only 5 years of data are available, no more than 45 independent model configurations should be tried. For that number of trials, the expected maximum SR IS is 1, whereas the expected SR OOS is 0. After trying only 7 independent strategy configurations, the expected maximum SR IS is 1 for a 2-year long backtest, while the expected SR OOS is 0. The implication is that a backtest which does not report the number of trials N used to identify the selected configuration makes it impossible to assess the risk of overfitting.

generates a random sequence of trades) on a sample of T observations. Overfitting will be difficult, because $N = 2$. Let's say that we make the model more complex, by adding 4 more parameters so that the total number of parameters becomes 5, i.e. $N = 2^5 = 32$. Having 32 independent sequences of random trades greatly increases the possibility of overfitting.

While a greater N makes overfitting easier, it makes perfectly fitting harder. Modern supercomputers can only perform around 2^{50} raw computations per second, or less than 2^{58} raw computations per year. Even if a trial could be reduced to a raw computation, searching $N = 2^{100}$ will take us 2^{42} supercomputer-years of computation (assuming a 1 Pflop/s system, capable of 10^{15} floating-point operations per second). Hence, a skill-less brute force search is certainly impossible. While it is hard to perfectly fit a complex skill-less strategy, Theorem 1 shows that there is no need for that. Without perfectly fitting a strategy, or making it over-complex, a researcher can achieve high Sharpe ratios. A relatively simple strategy with just 7 binomial independent parameters offers $N = 2^7 = 128$ trials, with an expected maximum Sharpe ratio above 2.6 (see Figure 1).

We suspect, however, that backtested strategies that significantly beat the market typically rely on some combination of valid insight, boosted by some degree of overfitting. Since believing in such an artificially enhanced high performance strategy will often also lead to over-leveraging, such overfitting is still very damaging. Most Technical Analysis strategies rely on filters, which are sets of conditions that trigger trading actions, like the random switches exemplified earlier. Accordingly, extra caution is warranted to guard against overfitting in using Technical Analysis strategies, as well as in complex non-parametric modeling tools, such as Neural Networks and Kernel Estimators.

Here is a key concept that investors generally miss:

> *A researcher that does not report the number of trials N used to identify the selected backtest configuration makes it impossible to assess the risk of overfitting.*

Because N is almost never reported, the magnitude of overfitting in published backtests is unknown. It is not hard to overfit a backtest (indeed, the previous theorem shows that it is hard *not to*), so we suspect that a large proportion of backtests published in academic journals may be misleading. The situation is not likely to be better among practitioners.

In our experience, overfitting is pathological within the financial industry, where proprietary and commercial software is developed to estimate the

combination of parameters that best fits (or more precisely, overfits) the data. These tools allow the user to add filters without ever reporting how such additions increase the probability of backtest overfitting. Institutional players are not immune to this pitfall. Large mutual fund groups typically discontinue and replace poorly performing funds, introducing *survivorship* and selection bias. While the motivation of this practice may be entirely innocent, the effect is the same as that of hiding experiments and inflating expectations.

We are not implying that those technical analysts, quantitative researchers or fund managers are "snake oil salesmen." Most likely they genuinely believe that the backtested results are legitimate, or that adjusted fund offerings better represent future performance. Hedge fund managers are often unaware that most backtests presented to them by researchers and analysts may be useless, and so they unknowingly package faulty investment propositions into products. One goal of this paper is to make investors, practitioners and academics aware of the futility of considering backtest without controlling for the probability of overfitting.

5   OVERFITTING IN ABSENCE OF COMPENSATION EFFECTS

Regardless of how realistic the prior being tested is, there is always a combination of parameters that is optimal. In fact, even if the prior is false, the researcher is very likely to identify a combination of parameters that happens to deliver an outstanding performance IS. But because the prior is false, OOS performance will almost certainly underperform the backtest's results. As we have described, this phenomenon, by which IS results tend to outperform the OOS results, is called overfitting. It occurs because a sufficiently large number of parameters are able to target specific data points - say by chance buying just before a rally and shorting a position just before a sell-off – rather than triggering trades according to the prior.

To illustrate this point, suppose we generate N Gaussian random walks by drawing from a Standard Normal distribution, each walk having a size T. Each performance path m_τ can be obtained as a cumulative sum of Gaussian draws

$$\Delta m_\tau = \mu + \sigma \varepsilon_\tau \tag{5.1}$$

where the *random shocks* ε_τ are IID distributed $\varepsilon_\tau \sim Z, \tau = 1, \ldots, T$. Suppose that each path has been generated by a particular combination of

14

parameters, backtested by a researcher. Without loss of generality, assume that $\mu = 0, \sigma = 1$ and $T = 1000$, covering a period of one year (with about 4 observations per trading day). We divide these paths into two disjoint samples of equal size 500, and call the first one IS and the second one OOS.

At the moment of choosing a particular parameter combination as optimal, the researcher had access to the IS series, not the OOS. For each model configuration, we may compute the Sharpe ratio of the series IS, and compare it with the Sharpe ratio of the series OOS. Figure 3 shows the resulting scatter plot. The p-values associated with the intercept and the IS performance (SR *a priori*) are respectively 0.6261 and 0.7469.

The problem of overfitting arises when the researcher uses the IS performance (backtest) to choose a particular model configuration, with the expectation that configurations that performed well in past will continue to do so in future. This would be a correct assumption if the parameter configurations were associated with a truthful prior, but this is clearly not the case of the simulation above, which is the result of Gaussian random walks without trend ($\mu = 0$).

Figure 4 shows what happens when we select the model configuration associated with the random walk with highest Sharpe ratio IS. The performance of the first half was optimized IS, and the performance of the second half is what the investor receives OOS. The good news is that under these conditions, there is no reason to expect overfitting to induce negative performance. This is illustrated in Figure 5, which shows how the optimization causes the expected performance IS to range between 1.2 and 2.6, while the OOS performance will range between -1.5 and 1.5 (i.e., around μ, which in this case is zero). The p-values associated with the intercept and the IS performance (SR *a priori*) are respectively 0.2146 and 0.2131. Selecting an optimal model IS had no bearing on the performance OOS, which simply equals the zero mean of the process. A positive mean ($\mu > 0$) would lead to positive expected performance OOS, but such performance would nevertheless be inferior to the one observed IS.

6 OVERFITTING IN PRESENCE OF COMPENSATION EFFECTS

Multiple causes create compensation effects in practice, such as overcrowded investment opportunities, major corrections, economic cycles, reversal of financial flows, structural breaks, bubble bursts, etc. Optimizing a strategy's parameters (i.e., choosing the model configuration that maximizes the strat-
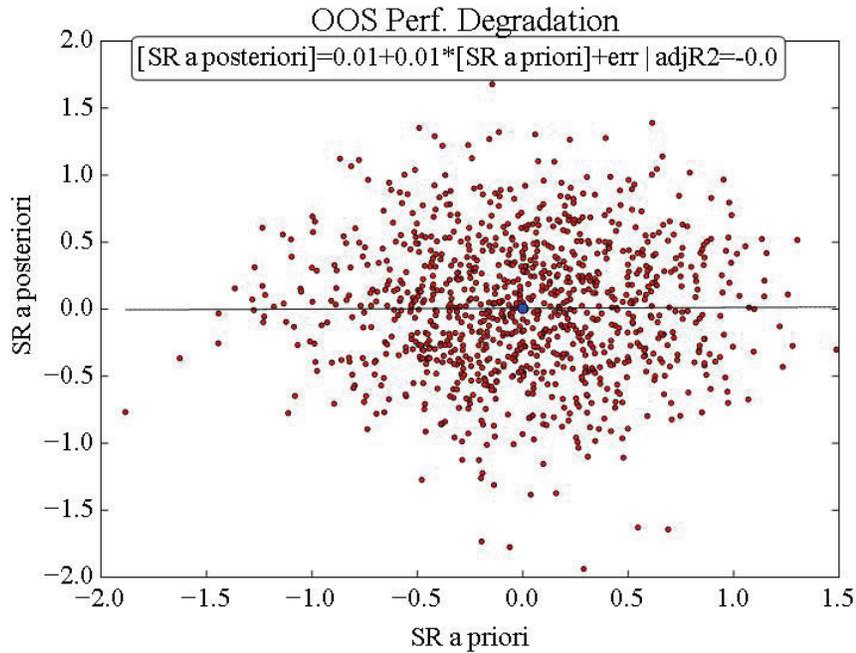
Figure 3: Performance IS vs. OOS before introducing strategy selection

Figure 3 shows the relation between SR IS (x-axis) and SR OOS (y-axis), for $\mu = 0, \sigma = 1, N = 1000, T = 1000$. Because the process follows a random walk, the scatter plot has a circular shape centered in the point (0,0). This illustrates the fact that, in absence of compensation effects, overfitting IS performance (x-axis) has no bearing on the OOS performance (y-axis), which remains around zero.
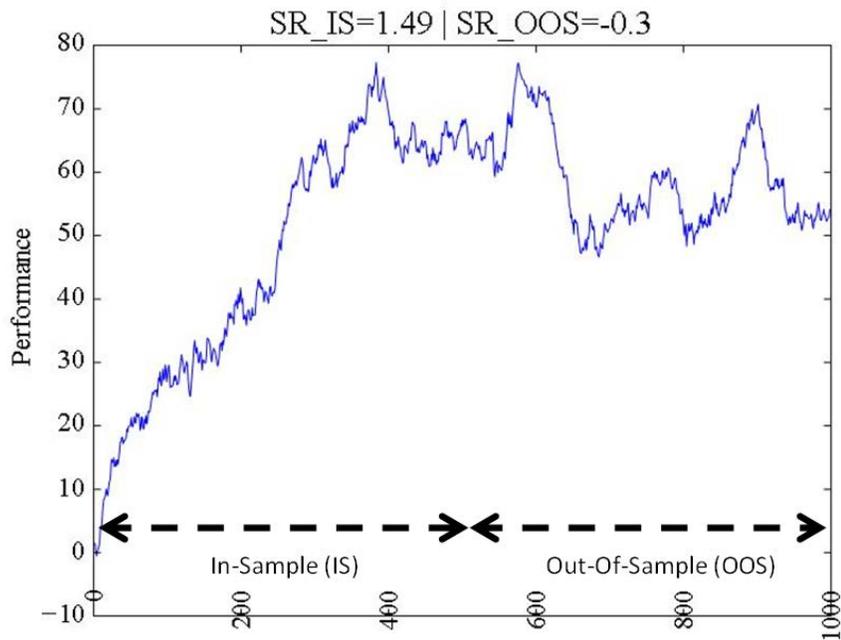
Figure 4: Performance IS vs. performance OOS for one path, after introducing strategy selection

Figure 4 provides a graphical representation of what happens when we select the random walk with highest SR IS. The performance of the first half was optimized IS, and the performance of the second half is what the investor receives OOS. The good news is, in the absence of memory, there is no reason to expect overfitting to induce negative performance.

17
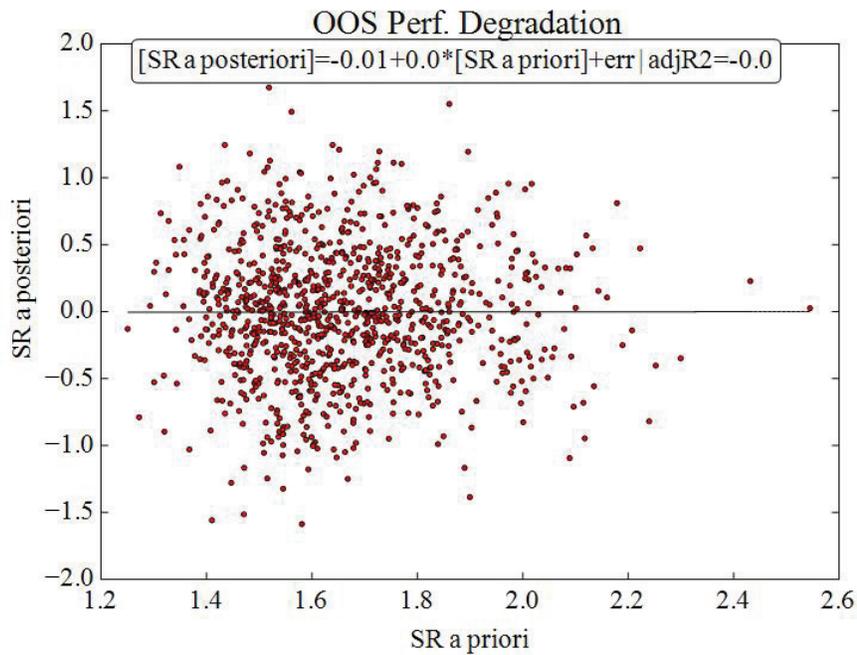
Figure 5: Performance degradation after introducing strategy selection, in absence of compensation effects

Figure 5 illustrates what happens once we add a "model selection" procedure. Now the SR IS ranges from 1.2 to 2.6, and it is centered around 1.7. Although the backtest for the selected model generates the expectation of a 1.7 SR, the expected SR OOS is unchanged and lies around 0.

egy's performance IS) does not necessarily lead to improved performance (compared to not optimizing) OOS, yet again leading to overfitting.

In some instances, when the strategy's performance series lacks memory, overfitting leads to no improvement in performance OOS. However, the presence of memory in a strategy's performance series induces a compensation effect, which increases the chances for that strategy to be selected IS, only to underperform the rest OOS. Under those circumstances, IS backtest optimization is in fact detrimental to OOS performance.[4]

6.1 GLOBAL CONSTRAINT

Unfortunately, overfitting rarely has the neutral implications discussed in the previous section. Our previous example was purposely chosen to exhibit a globally unconditional behavior. As a result, the OOS data had no memory of what occurred IS. Centering each path to match a mean μ removes one degree of freedom.

$$\overline{\Delta m_\tau} = \Delta m_\tau + \mu - \frac{1}{T} \sum_{\tau=1}^{T} \Delta m_\tau \qquad (6.1)$$

We may re-run the same Monte Carlo experiment as before, this time on the re-centered variables $\overline{\Delta m_\tau}$. Somewhat scarily, adding this single global constraint causes the OOS performance to be negative, even though the underlying process was trendless. Moreover, a strongly negative linear relation between performance IS and OOS arises, indicating that the more we optimize IS, the worse is OOS performance. Figure 6 displays this disturbing pattern. The p-values associated with the intercept and the IS performance (SR *a priori*) are respectively 0.5005 and 0, indicating that the negative linear relation between IS and OOS Sharpe ratios is statistically significant.

The following proposition is proven in the Appendix.

Proposition 6.1. *Given two alternative configurations (A and B) of the same model, where $\sigma_{IS}^A = \sigma_{OOS}^A = \sigma_{IS}^B = \sigma_{OOS}^B$ imposing a global constraint $\mu^A = \mu^B$ implies that*

$$SR_{IS}^A > SR_{IS}^B \Leftrightarrow SR_{OOS}^A < SR_{OOS}^B \qquad (6.2)$$

Recentering a series is one way to introduce memory into a process, because some data points will now compensate for the extreme outcomes

---

[4]Bailey et al. [1] propose a method to determine the degree to which a particular backtest may have been compromised by the risk of overfitting.
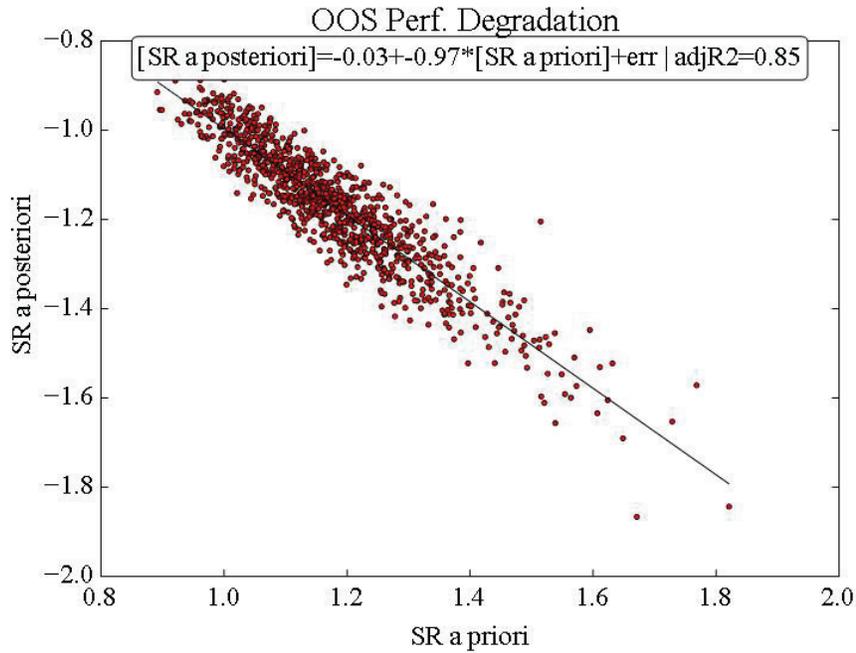
Figure 6: Performance degradation as a result of strategy selection under compensation effects (global constraint)

Adding a single global constraint causes the OOS performance to be negative, even though the underlying process was trendless. Also, a strongly negative linear relation between performance IS and OOS arises, indicating that the more we optimize IS, the worse will be the OOS performance of the strategy.

from other data points. By optimizing a backtest, the researcher selects a model configuration that spuriously works well IS, and consequently is likely to generate losses OOS.

6.2 SERIAL DEPENDENCE

But imposing a global constraint is not the only situation in which overfitting actually is detrimental. To cite another (less restrictive) example, the same effect happens if the performance series is serially conditioned, such as a first-order autoregressive process.

$$\Delta m_\tau = (1 - \varphi)\mu + (\varphi - 1)\varphi m_{\tau-1} + \sigma\varepsilon_\tau \tag{6.3}$$

or analogously,

$$m_\tau = (1 - \varphi)\mu + \varphi m_{\tau-1} + \sigma\varepsilon_\tau \tag{6.4}$$

where the random shocks are again are IID distributed as $\varepsilon_\tau \sim Z$. The following proposition is proven in the Appendix. The number of observations that it takes for a process to reduce its divergence from the long-run equilibrium by half is known as the *half-life period*, or simply *half-life* (a familiar physical concept introduced by Ernest Rutherford in 1907).

Proposition 6.2. *The half-life period of a first-order autoregressive process with autoregressive coefficient $\varphi \in (0, 1)$ occurs at*

$$\tau = -\frac{\ln[2]}{\ln[\varphi]}. \tag{6.5}$$

For example, if $\varphi = 0.995$, it takes about 138 observations to retrace half of the deviation from the equilibrium. This introduces another form of compensation effect, just as we saw in the case of a global constraint. If we re-run the previous Monte Carlo experiment, this time for the autoregressive process with $\mu = 0, \sigma = 1, \varphi = 0.995$, and plot the pairs of performance IS vs. OOS, we obtain Figure 7.

The p-values associated with the intercept and the IS performance (SR *a priori*) are respectively 0.4513 and 0, confirming that the negative linear relation between IS and OOS Sharpe ratios is again statistically significant. Such serial correlation is a well-known statistical feature, present in the performance of most hedge fund strategies. Proposition 6.3 is proved in the Appendix.
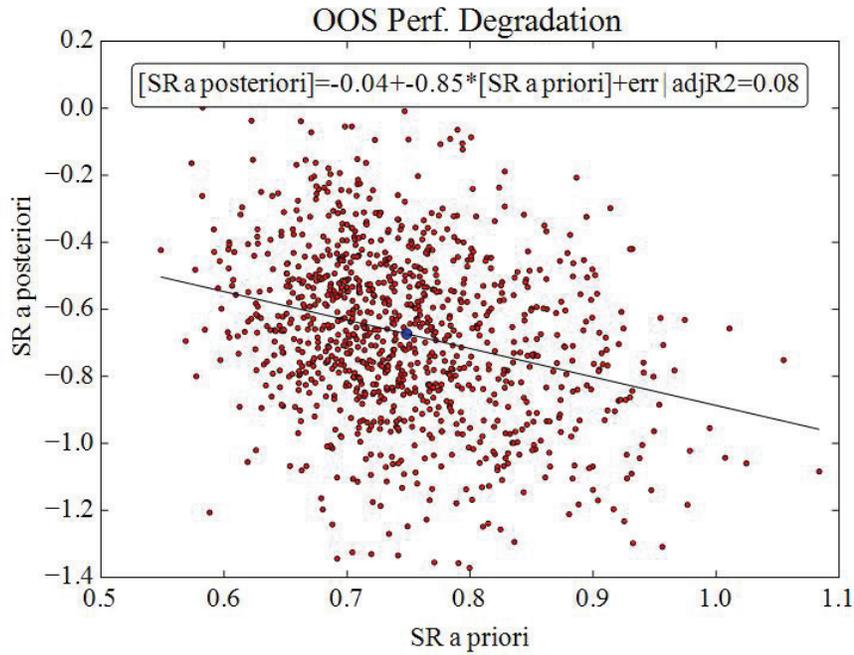
Figure 7: Performance degradation as a result of strategy selection under compensation effects (first-order serial correlation)

Serially-correlated performance introduces another form of compensation effects, just as we saw in the case of a global constraint. For example, if $\varphi = 0.995$, it takes about 138 observations to recover half of the deviation from the equilibrium. We have re-run the previous Monte Carlo experiment, this time on an autoregressive process with $\mu = 0, \sigma = 1, \varphi = 0.995$, and plotted the pairs of performance IS vs. OOS.

Proposition 6.3. *Given two alternative configurations (A and B) of the same model, where $\sigma_{IS}^A = \sigma_{OOS}^A = \sigma_{IS}^B = \sigma_{OOS}^B$ and the performance series follows the same first-order autoregressive stationary process,*

$$SR_{IS}^A > SR_{IS}^B \Leftrightarrow SR_{OOS}^A < SR_{OOS}^B \tag{6.6}$$

Proposition 6.3 reaches the same conclusion as Proposition 6.1 (a compensation effect), without requiring a global constraint.

7 IS BACKTEST OVERFITTING A FRAUD?

Consider an investment manager who e-mails his stock market forecast for the next month to $2^n x$ prospective investors, where x and n are positive integers. To half of them he predicts that markets will go up, and to the other half that markets will go down. After the month passes, he drops from his list the names to which he sent the incorrect forecast, and resend a new forecast to the remaining $2^{n-1}x$ names. He repeats the same procedure n times, after which only x names remain. These x investors have witnessed n consecutive infallible forecasts, and may be extremely tempted to give this investment manager all of their savings. Of course, this is a fraudulent scheme based on random screening: The investment manager is hiding that for every one of the x successful witness, he has tried 2^n unsuccessful ones (see Harris [8, p. 473] for a similar example).

To avoid falling for this psychologically compelling fraud, a potential investor needs to consider the economic cost associated with manufacturing the successful experiments, and require the investment manager to produce a number n for which the scheme is uneconomic. One caveat is, even if n is too large for a skill-less investment manager, it may be too low for a mediocre investment manager who uses this scheme to inflate his skills.

Not reporting the number of trials (N) involved in identifying a successful backtest is a similar kind of fraud. The investment manager only publicizes the model that works, but says nothing about all the failed attempts, which as we have seen can greatly increase the probability of backtest overfitting.

An analogous situation occurs in medical research, where drugs are tested by treating hundreds or thuosands of patients, however only the best outcomes are publicized. The reality is that the selected outcomes may have healed despite of (rather than thanks to) the treatment, or due to a placebo effect (recall Theorem 1). Such behavior is unscientific—not to mention dangerous and expensive—and has led to the launch of the *alltrials.net* project, which demands that all results (positive and negative) for every experiment

are made publicly available. A step forward in this direction is the recent announcement by Johnson and Johnson that it will plans to open all of its clinical test results to the public [14]. For related discussion of reproducibility in the context of mathematical computing, see Stodden, et al. [25].

Hiding trials appears to be standard procedure in financial research and financial journals. As an aggravating factor, we know from Section 6 that backtest overfitting typically has a detrimental effect on future performance, due to the compensation effects present in financial series. Indeed, the customary disclaimer *"past performance is not an indicator of future results"* is too optimistic in the context for backtest overfitting. When investment advisers do not control for backtest overfitting, good backtest performance is an indicator of negative future results.

8   A PRACTICAL APPLICATION

Institutional asset managers follow certain investment procedures on a regular basis, such as rebalancing the duration of a fixed income portfolio (PIMCO), rolling holdings on commodities (Goldman Sachs, AIG, JP Morgan, Morgan Stanley), investing or divesting as new funds flow at the end of the month (Fidelity, BlackRock), participating in the regular U.S. Treasury Auctions (all major investment banks), de-levering in anticipation of payroll, FOMC or GDP releases, tax-driven effects around the end of the year and mid-April, positioning for electoral cycles, etc. There is a large number of instances where asset managers will engage in somewhat predictable actions on a regular basis. It should come as no surprise that a very popular investment strategy among hedge funds is to profit from such seasonal effects.

For example, a type of question often asked by hedge fund managers follow the form: "Is there a time interval every _____ when I would have made money on a regular basis?" You may replace the blank space with a word like day, week, month, quarter, auction, nonfarm payroll (NFP) release, European Central Bank (ECB) announcement, presidential year, ... The variations are as abundant as they are inventive. Doyle and Chen [4] study the "weekday effect" and conclude that it appears to "wander".

The problem with this line of questioning is that there is always a time interval that is arbitrarily "optimal," regardless of the cause. The answer to one such questions is the title of a very popular investment classic, "Do not sell stocks on Monday," by Hirsch [12]. The same author wrote an almanac for stock traders that reached its 45th edition in 2012, and is also a pro-

ponent of the *"Santa Claus Rally,"* the quadrennial political/stock market cycle, and investing during the *"Best Six Consecutive Months"* of the year, November through April. While these findings may indeed be caused by some underlying seasonal effect, it is easy to demonstrate that any random data contains similar patterns. The discovery of a pattern IS typically has no bearing OOS, yet again as a result of overfitting. Running such experiments without controlling for the probability of backtest overfitting will lead the researcher to spurious claims. OOS performance will disappoint, and the reason will not be that "the market has found out the seasonal effect and arbitraged away the strategy's profits." Rather, the effect was never there; instead it was just a random pattern that gave rise to an overfitted trading rule. We will illustrate this point with an example.

Example 8.1. 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) which triggers an exit for that month's position. For example, we could explore all nodes that span the set $\{1, \ldots, 22\}$ for *Entry_day*, the set $\{1, \ldots, 20\}$ for *Holding_period*, the set $\{0, \ldots, 10\}$ for *Stop_loss*, and $\{-1, 1\}$ for *Side*. The parameter 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, we generated a time series of 1000 daily prices (about 4 years), following a random walk. Figure 8 plots the random series, as well as the performance associated with the optimal parameter combination: Entry_day = 11, Holding_period = 4, Stop_loss = -1 and Side = 1. The 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 [5]. Several studies in the practitioners and

---

[5]The Probabilistic Sharpe Ratio (or PSR) is an extension to the SR. Non-Normality increases the error of the variance estimator, and PSR takes that into consideration when determining whether a SR estimate is statistically significant. See Bailey and López de Prado [2] for details.
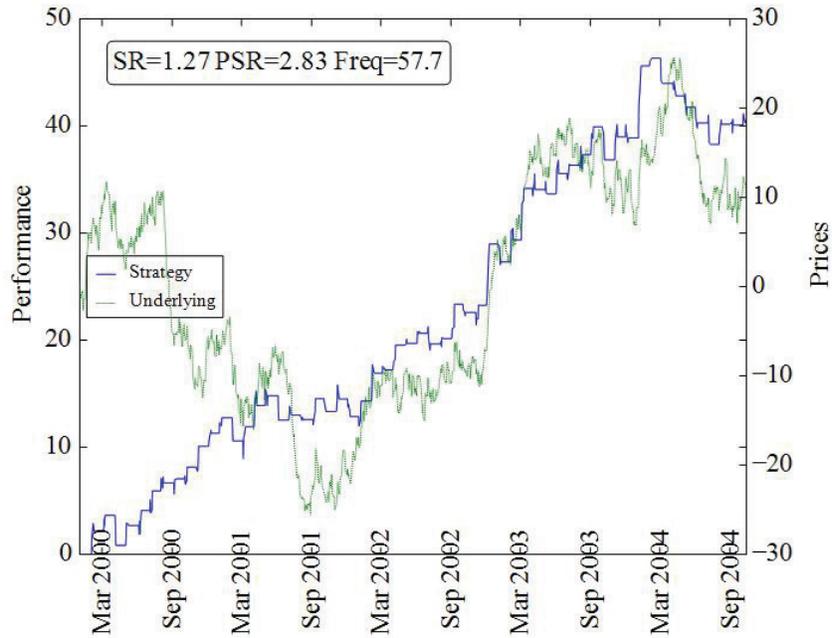
Figure 8: Backtested performance of a seasonal strategy (example 1)

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 plausible seasonal strategy with a SR of 1.27 despite the fact that no true seasonal effect exists.

academic literature report similar results, which are conveniently justified with some ex-post explanation ("the posterior gives rise to a prior"). What this analysis misses is an evaluation of the probability that this backtest has been overfit to the data, which is the subject of Bailey et al. [1].

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 Section 10.5 for the implementation of this experiment in the 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 highly suspicious of backtests and of researchers who fail to report the number of trials attempted.

9 CONCLUSIONS

While the literature on regression overfitting is extensive, we believe that this is the first study to discuss the issue of overfitting in the context of investment simulations (backtests), and its negative effect on OOS performance. On the subject of regression overfitting, the great Enrico Fermi once remarked (Mayer et al. [20]):

> I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

The same principle applies to backtesting, with some interesting peculiarities. We have shown that backtest overfitting is difficult indeed to avoid. Any perseverant researcher will always be able to find a backtest with a desired Sharpe ratio, regardless of the sample length requested. Model complexity is only one way that backtest overfitting is facilitated. Given that most published backtests do not report the number of trials attempted, many of them may be overfitted. In that case, if an investor allocates them capital, performance will vary: *It will be around zero if the process has no memory, but it may be significantly negative if the process has memory.* The standard warning that "past performance is not an indicator of future results" understates the risks associated with investing on overfit backtests. When financial advisors do not control for overfitting, positive backtested performance will often be followed by negative investment results.

27

We have derived the expected maximum Sharpe ratio as a function of the number of trials (N) and sample length. This has allowed us to determine the *Minimum Backtest Length* (MinBTL) needed to avoid selecting a strategy with a given IS Sharpe ratio among N trials with an expected OOS Sharpe ratio of zero. Our conclusion is that, the more trials a financial analyst executes, the greater should be the IS Sharpe ratio demanded by the potential investor.

We strongly suspect that such backtest overfitting is a large part of the reason why so many algorithmic or systematic hedge funds do not live up to the elevated expectations generated by their managers.

We would feel sufficiently rewarded in our efforts if at least this paper succeeded in drawing the attention of the mathematical community regarding the widespread proliferation of journal publications, many of them claiming profitable investment strategies on the sole basis of in-sample performance. This is understandable in business circles, but a higher standard is and should be expected from an academic forum.

A depressing parallel can be drawn between today's financial academic research and the situation denounced by economist and Nobel Laureate Wassily Leontief writing in *Science* (see Leontief [1982]):

> A dismal performance
>
> ... "What economists revealed most clearly was the extent to which their profession lags intellectually." This editorial comment by the leading economic weekly (on the 1981 annual proceedings of the American Economic Association) says, essentially, that the "king is naked." But no one taking part in the elaborate and solemn procession of contemporary U.S. academic economics seems to know it, and those who do don't dare speak up
>
> [. . .]
>
> [E]conometricians fit algebraic functions of all possible shapes to essentially the same sets of data without being able to advance, in any perceptible way, a systematic understanding of the structure and the operations of a real economic system
>
> [. . .]
>
> That state is likely to be maintained as long as tenured members of leading economics departments continue to exercise tight control over the training, promotion, and research activities of their younger faculty members and, by means of peer review, of the senior members as well.

We hope that our distinguished colleagues will follow this humble attempt with ever deeper and more convincing analysis. We did not write this paper to settle a discussion. On the contrary, our wish is to ignite a

dialogue among mathematicians, and a reflection among investors and regulators. We should do well also to heed Newton's comment after he lost heavily in the South seas bubble, see [21].

> For those who had realized big losses or gains, the mania redistributed wealth. The largest honest fortune was made by Thomas Guy, a stationer turned philanthropist, who owned £54,000 of South Sea stock in April 1720 and sold it over the following six weeks for £234,000. Sir Isaac Newton, scientist, master of the mint, and a certifiably rational man, fared less well. He sold his £7,000 of stock in April for a profit of 100 percent. But something induced him to reenter the market at the top, and he lost £20,000. "I can calculate the motions of the heavenly bodies," he said,"but not the madness of people."

10 APPENDICES

10.1 PROOF OF PROPOSITION 2.1

Embrechts et al. [5, 138–147] show that the maximum value (or last order statistic) in a sample of independent random variables following an exponential distribution converges asymptotically to a Gumbel distribution. As a particular case, the Gumbel distribution covers the Maximum Domain of Attraction of the Gaussian distribution, and therefore it can be used to estimate the expected value of the maximum of several independent random Gaussian variables.

To see how, suppose a sample of IID random variables, $z_n \sim Z, n = 1, \ldots, N$, where Z is the CDF of the Standard Normal distribution. To derive an approximation for the sample maximum, $\max_n = \max\{z_n\}$, we apply the Fisher-Tippet-Gnedenko theorem to the Gaussian distribution, and obtain that

$$\lim_{N \to \infty} Prob\left[\frac{\max_N - \alpha}{\beta} \leq x\right] = G[x] \tag{10.1}$$

where

- $G[x] = e^{-e^{-x}}$ is the CDF for the Standard Gumbel distribution.

- $\alpha = Z^{-1}\left[1 - \frac{1}{N}\right]$, $\beta = Z^{-1}\left[1 - \frac{1}{N}e^{-1}\right] - \alpha$, and Z^{-1} corresponds to the inverse of the Standard Normal's CDF.

The normalizing constants (α, β) are derived in Resnick [22] and Embrechts et al. [5]. The limit of the expectation of the normalized maxima

from a distribution in the Gumbel Maximum Domain of Attraction (see Proposition 2.1(iii) in Resnick [22]) is

$$\lim_{N \to \infty} E\left[\frac{\max_N - \alpha}{\beta}\right] = \gamma \tag{10.2}$$

where γ is the Euler-Mascheroni constant, $\gamma \approx 0.5772156649\ldots$ Hence, for N sufficiently large, the mean of the sample maximum of standard normally distributed random variables can be approximated by

$$E[\max_N] \approx \alpha + \gamma\beta = (1 - \gamma)Z^{-1}\left[1 - \frac{1}{N}\right] + \gamma Z^{-1}\left[1 - \frac{1}{N}e^{-1}\right] \tag{10.3}$$

where $N > 1$. Q.E.D.

10.2 PROOF OF PROPOSITION 6.1

Suppose two random samples (A and B) of the same process $\{\Delta m_\tau\}$, where A and B are of equal size, and have means and standard deviations $\mu^A, \mu^B, \sigma^A, \sigma^B$. A fraction δ of each sample is called IS, and the remainder is called OOS, where for simplicity we have assumed that $\sigma_{IS}^A = \sigma_{OOS}^A = \sigma_{IS}^B = \sigma_{OOS}^B$. We would like to understand the implications of a global constraint $\mu^A = \mu^B$.

First, we note that $\mu^A = \delta\mu_{IS}^A + (1-\delta)\mu_{OOS}^A$ and $\mu^B = \delta\mu_{IS}^B + (1-\delta)\mu_{OOS}^B$. Then, $\mu_{IS}^A > \mu_{OOS}^A \Leftrightarrow \mu_{IS}^A > \mu^A \Leftrightarrow \mu_{OOS}^A < \mu^A$. Likewise, $\mu_{IS}^B > \mu_{OOS}^B \Leftrightarrow \mu_{IS}^B > \mu^B \Leftrightarrow \mu_{OOS}^B < \mu^B$.

Second, because of the global constraint $\mu^A = \mu^B$, $\mu_{IS}^A + \frac{(1-\delta)}{\delta}\mu_{OOS}^A = \mu_{IS}^B + \frac{(1-\delta)}{\delta}\mu_{OOS}^B$ and $\mu_{IS}^A - \mu_{IS}^B = \frac{(1-\delta)}{\delta}(\mu_{OOS}^B - \mu_{OOS}^A)$. Then, $\mu_{IS}^A > \mu_{IS}^B \Leftrightarrow \mu_{OOS}^A < \mu_{OOS}^B$. We can divide this expression by $\sigma_{IS}^A > 0$, with the implication is that

$$SR_{IS}^A > SR_{IS}^B \Leftrightarrow SR_{OOS}^A < SR_{OOS}^B \tag{10.4}$$

where we have denoted $SR_{IS}^A = \frac{\mu_{IS}^A}{\sigma_{IS}^A}$, etc. Note that we did not have to assume that Δm_τ is IID, thanks to our assumption of equal standard deviations. The same conclusion can be reached without assuming equality of standard deviations, however the proof would be longer but no more revealing (the point of this proposition is the implication of global constraints). Q.E.D.

10.3 PROOF OF PROPOSITION 6.2

This proposition computes the half-life of a first-order autoregressive process. Suppose a random variable m_τ that takes values of a sequence of observations $\tau \in \{1, \ldots, \infty\}$, where

$$m_\tau = (1 - \varphi)\mu + \varphi m_{\tau-1} + \sigma \varepsilon_\tau \tag{10.5}$$

such that the random shocks are IID distributed as $\varepsilon_\tau \sim N(0, 1)$.

$$\lim_{\tau \to \infty} E_0[m_\tau] = \mu$$

if and only if $\varphi \in (-1, 1)$. In particular, from Bailey and López de Prado [3] we know that the expected value of this process at a particular observation τ is

$$E_0[m_\tau] = \mu(1 - \varphi^\tau) + \varphi^\tau m_0 \tag{10.6}$$

Suppose that the process is initialized or reset at some value $m_0 \neq \mu$. We ask the question, how many observations must pass before

$$E_0[m_\tau] = \frac{\mu + m_0}{2}? \tag{10.7}$$

Inserting Eq. (10.7) into Eq. (10.6), and solving for τ, we obtain

$$\tau = \frac{\ln[2]}{\ln[\varphi]} \tag{10.8}$$

which implies the additional constraint that $\varphi \in (0, 1)$.   Q.E.D.

10.4 PROOF TO PROPOSITION 6.3

Suppose that we draw two samples (A and B) of a first-order autoregressive process, and generate to subsamples of each. The first subsample is called IS, and it is comprised of $\tau = 1, \ldots, \delta T$, and the second subsample is called OOS, as it is comprised of $\tau = \delta T+1, \ldots, T$ with $\delta \in (0, 1)$, and T an integer multiple of δ. For simplicity, let us assume that $\sigma_{IS}^A = \sigma_{OOS}^A = \sigma_{IS}^B = \sigma_{OOS}^B$. From Proposition 6.2, Eq. (10.5) we obtain

$$E_{\delta T}[m_T] - m_{\delta T} = (1 - \varphi^T)(\mu - m_{\delta T}) \tag{10.9}$$

Because $1 - \varphi^T > 0$, $sigma_{IS}^A = \sigma_{IS}^B$, $SR_{IS}^A > SR_{IS}^B \Leftrightarrow m_{\delta T}^A > m_{\delta T}^B$. This means that the OOS of A begins with a seed which is greater than the seed that initializes the OOS of B. Therefore, $m_{\delta T}^A > m_{\delta T}^B \Leftrightarrow E_{\delta T}[m_T^A] - m_{\delta T}^A < E_{\delta T}[m_T^B] - m_{\delta T}^B$. Because $\sigma_{IS}^B = \sigma_{OOS}^B$, we conclude that

$$SR_{IS}^A > SR_{IS}^B \Leftrightarrow SR_{OOS}^A < SR_{OOS}^B \tag{10.10}$$

$$\text{Q.E.D.}$$

10.5   REPRODUCING THE RESULTS IN SECTION 9

Python code implementing the experiment described in Section 9 is lodged at `http://www.quantresearch.info/Software.htm` and at `http://www.financial-math.org/software/`.

References

[1] D. Bailey, J. Borwein, M. López de Prado and J. Zhu, "The probability of backtest overfitting," working paper, 2013, http://ssrn.com/abstract=2326253.

[2] D. Bailey and M. López de Prado, "The Sharpe Ratio efficient frontier," *Journal of Risk*, 15(2) (2012), 3–44, working paper, http://ssrn.com/abstract=1821643.

[3] D. Bailey and M. López de Prado, "Drawdown-based stop-outs and the triple penance rule," working paper, 2013, http://ssrn.com/abstract=2201302.

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

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

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

[7] J. Hadar and W. Russell, "Rules for ordering uncertain prospects," *American Economic Review*, 59 (1969), 25–34.

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

[9] C. Harvey and Y. Liu, "Backtesting," working paper, 2013, http://ssrn.com/abstract=2345489.

[10] C. Harvey, Y. Liu and H. Zhu, "...And the cross-section of expected returns," working paper, 2013, http://ssrn.com/abstract=2249314.

[11] D. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Science*, 44 (2004), 1–12.

[12] Y. Hirsch, "Don't sell stocks on Monday," *Penguin Books*, 1st Edition, 1987.

[13] J. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, 2(8) (Aug 2005).

[14] H. Krumholz, "Give the data to the people," *New York Times*, 2 Feb 2014, http://www.nytimes.com/2014/02/03/opinion/give-the-data-to-the-people.html.

[15] D. Leinweber and K. Sisk, "Event driven trading and the 'new news'," *Journal of Portfolio Management*, 38(1) (2011), 110–124.

[16] W. Leontief, "Academic economics," *Science Magazine*, 9 Jul 1982, 104–107.

[17] A. Lo, "The statistics of Sharpe Ratios," *Financial Analysts Journal*, (58)4 (Jul/Aug 2002), http://ssrn.com/abstract=377260.

[18] M. López de Prado and A. Peijan, "Measuring the loss potential of hedge fund strategies," *Journal of Alternative Investments*, 7(1) (2004), 7–31, http://ssrn.com/abstract=641702.

[19] M. López de Prado and M. Foreman, "A mixture of Gaussians approach to mathematical portfolio oversight: The EF3M algorithm," working paper, RCC at Harvard University, http://ssrn.com/abstract=1931734.

[20] J. Mayer, K. Khairy and J. Howard, "Drawing an elephant with four complex parameters," *American Journal of Physics*, 78(6) (2010).

[21] C. Reed, "The damn'd south sea," *Harvard Magazine*, May-Jun 1999.

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

[23] J. Romano and M. Wolf, "Stepwise multiple testing as formalized data snooping," *Econometrica*, 73(4) (2005), 1273–1282.

[24] F. Schorfheide and K. Wolpin, "On the use of holdout samples for model selection," *American Economic Review*, 102(3) (2012), 477–481.

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

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

[27] W. White, "A reality check for data snooping," *Econometrica*, 68(5) (2000), 1097–1126.

[28] S. Weiss 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.

[29] A. Wiles, "Financial greed threatens the good name of maths," *The Times*, 04 Oct 2013, `http://www.thetimes.co.uk/tto/education/article3886043.ece`.