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

David H. Bailey [α]

Jonathan M. Borwein [β]

Marcos López de Prado [γ]

Qiji Jim Zhu [δ]

First version: September 2013
This version: 7 October 2013

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

[β] Jonathan M. Borwein is Laureate Professor of Mathematics at University of Newcastle, Callaghan NSW 2308, Australia, and a Fellow of the Royal Society of Canada, the Australian Academy of Science and of the AAAS. E-mail: jonathan.borwein@newcastle.edu.au. URL: www.carma.newcastle.edu.au/jon

[γ] Marcos López de Prado is Head of Quantitative Trading & Research at Hess Energy Trading Company, New York, NY 10036, and a Research Affiliate at Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. E-mail: lopezdeprado@lbl.gov. URL: www.QuantResearch.info

[δ] Qiji Jim Zhu is Professor of Mathematics at Western Michigan University, Kalamazoo, MI 49008. E-mail: zhu@math-stat.wmich.edu. URL: http://homepages.wmich.edu/~zhu/

We are grateful to Tony Anagnostakis (Moore Capital), Marco Avellaneda (Courant Institute, NYU), Peter Carr (Morgan Stanley, NYU), Paul Embrechts (ETH Zürich), Matthew D. Foreman (University of California, Irvine), Ross Garon (SAC Capital), Attilio Meucci (Kepos Capital, NYU), Natalia Nolde (University of British Columbia and ETH Zürich) and Riccardo Rebonato (PIMCO, University of Oxford).

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.

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

ABSTRACT

Recent computational advances allow investment managers to search for profitable investment strategies. In many instances, that search 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 proposals.

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

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.

"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 results, 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 backtests' statistics, and allocate capital to the best performing.

As it relates 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 for profitable investment strategies. 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*. 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 momemtum" 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.

The machine learning literature has devoted significant effort to study the problem of overfitting. Their proposed methods typically are not applicable to investment strategy 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 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 do not typically 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 standard measure of performance in the presence of risk) above 2, based on a forecasting equation for which the AIC statistic (Akaike Information Criterion, a standard of regularization method) rejects the null hypothesis of overfitting with a 95% confidence level (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. The problem is, AIC's assessment did not take into account the hundreds of other trials that the researcher is hiding. For these reasons, 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 unanimously 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 [2013] find that there are hundreds of papers supposedly identifying hundreds of factors with explanatory power over future stocks returns. Their conclusion is that *"most claimed research findings are likely false."* Factor models are only the tip of the iceberg. We are certain that 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 [2011] 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 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, 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 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 founded. We encourage the reader to search the Internet for terms such as "stochastic oscillators," "Fibonacci ratios," "cycles," "Elliot wave," "Golden ratio," "Hindenberg Omen," "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. Even in 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 (SR) is a statistic that evaluates an investment manager or strategy's performance on the basis of a sample of past returns. Succinctly, 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. Suppose that a strategy's excess returns (or risk premiums), r_t, are *IID*

$$r_t \sim N(\mu, \sigma^2) \tag{1}$$

where *N* represents a Normal distribution with mean μ and variance σ^2. The purpose of the *Sharpe ratio* (SR) is to evaluate the skills of a particular strategy or investor.

$$SR = \frac{\mu}{\sigma} \tag{2}$$

Since μ, σ are usually unknown, the true value *SR* cannot be known for certain. The inevitable consequence is that Sharpe ratio calculations are likely to be the subject of substantial estimation errors.

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* returns distributed according to a Normal law with mean excess returns μ and with standard deviation σ. The number of returns per year is q. From Lo [2002], we know that the distribution of the estimated annualized Sharpe ratio $\widehat{SR}$ converges (for a sufficiently large *T*) to

$$\widehat{SR} \xrightarrow{a} N\left[SR, \frac{1 + \frac{1}{2}\frac{SR^2}{q}}{y}\right] \tag{3}$$

where $N[.]$ is the CDF of the Normal distribution, $SR = \frac{\mu}{\sigma}\sqrt{q}$ is the true annualized Sharpe ratio, $T = yq$ is the number of observations, and *y* is the number of years used to estimate $\widehat{SR}$.[1] Suppose that we would like to select the strategy with optimal $\widehat{SR}$ IS, based

[1] 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

on one year of observations. A risk we face is of 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*?

In Bailey and López de Prado [2012] we 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 paper we are concerned with overfitting prevention when comparing multiple strategies, not in evaluating the statistical significance of a single Sharpe ratio estimate. Next, we 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. (3), if $\mu = 0$ and $y = 1$, then $\widehat{SR} \overset{a}{\to} N[0,1]$. 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.

<u>*PROPOSITION 1*</u>*: Given a sample of IID random variables,* $x_n \sim \mathrm{Z}$, $n = 1, \ldots, N$, *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 N>1 as*

$$E[max_N] \approx (1-\gamma)\mathrm{Z}^{-1}\left[1-\frac{1}{N}\right] + \gamma\mathrm{Z}^{-1}\left[1-\frac{1}{N}e^{-1}\right] \tag{4}$$

where γ *(approx.. 0.5772156649) is the Euler-Mascheroni constant.*

An upper bound to Eq. (4) is $\sqrt{2Ln[N]}$. 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 model configurations, 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).

[FIGURE 1 HERE]

Proposition 1 has important implications. As long as the researcher tries more than one strategy configuration (N>1), 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.

---

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 [2012].

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

$$E[max_N] \approx y^{-1/2}\left((1-\gamma)\mathrm{Z}^{-1}\left[1-\frac{1}{N}\right]+\gamma\mathrm{Z}^{-1}\left[1-\frac{1}{N}e^{-1}\right]\right) \quad (5)$$

Eq. (5) 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. (5) for *y*, we reach the following statement.

THEOREM 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*

$$MinBTL \approx \left(\frac{(1-\gamma)\mathrm{Z}^{-1}\left[1-\frac{1}{N}\right]+\gamma\mathrm{Z}^{-1}\left[1-\frac{1}{N}e^{-1}\right]}{\overline{E[max_N]}}\right)^2 < \frac{2Ln[N]}{\overline{E[max_N]}^2} \quad (6)$$

Eq. (6) 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 1 assumed the *N* trials to be independent, which leads to a quite conservative estimate.

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), $MinBTL < \frac{2Ln[N]}{\overline{E[max_N]}^2}$.

[FIGURE 2 HERE]

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. [2013] 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 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. $N = 2^5 = 3\ 2$. 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.

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 over-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 published backtests 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 most 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

$$\Delta m_\tau = \mu + \sigma \varepsilon_\tau \tag{7}$$

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 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.

[FIGURE 3 HERE]

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.

[FIGURE 4 HERE]

[FIGURE 5 HERE]

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, bubbles' bursts, etc. Optimizing a strategy's parameters (i.e., choosing the model configuration that maximizes the strategy's performance IS) does not necessarily lead to improved performance OOS (compared to not optimizing), yet again leading ot 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.[2]

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.

[2] In Bailey et al. [2013] propose a method to determine the degree to which a particular backtest may have been compromised by the risk of overfitting.

$$\overline{\Delta m_\tau} = \Delta m_\tau + \mu - \frac{1}{T}\sum_{\tau=1}^{T} \Delta m_\tau \tag{8}$$

[FIGURE 6 HERE]

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 2: 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

$$SR_{IS}^{A} > SR_{IS}^{B} \Leftrightarrow SR_{OOS}^{A} < SR_{OOS}^{B} \tag{9}$$

Re-centering a series is one way to introduce memory into a process, because some data points will now compensate for the extreme outcomes from other data points. By optimizing a backtest, the researcher selects a model configuration that luckily works well IS, and consequently is likely to generate loses 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.

$$\Delta m_\tau = (1-\varphi)\mu + (\varphi - 1)\varphi m_{\tau-1} + \sigma\varepsilon_\tau \tag{10}$$

or analogously,

$$m_\tau = (1-\varphi)\mu + \varphi m_{\tau-1} + \sigma\varepsilon_\tau \tag{11}$$

where the random shocks areagain are IID distributed as $\varepsilon_\tau \sim Z$. The following proposition is proven in the Appendix.

PROPOSITION 3: The half-life period of a first-order autoregressive process with autoregressive coefficient $\varphi \in (0,1)$ occurs at

$$\tau = -\frac{Ln[2]}{Ln[\varphi]} \tag{12}$$

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). 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.

[FIGURE 7 HERE]

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 4 is proved in the Appendix.

PROPOSITION 4: 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,*

$$SR_{IS}^{A} > SR_{IS}^{B} \Leftrightarrow SR_{OOS}^{A} < SR_{OOS}^{B} \tag{13}$$

Proposition 4 reaches the same conclusion as Proposition 2 (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 resends 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 [2003, 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 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 and reprehensible in scientific research, 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. For related discussion of reproducibility in the context of mathematical computing, see Stodden, et a. (2013).

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, NFP release, ECB announcement, presidential year, … The variations are as abundant as they are inventive. Doyle and Chen [2009] 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 of such questions is the title of a very popular investment classic, "Do not sell stocks on Monday," by Hirsch [1987]. The same author wrote an almanac for stock traders that reached its 45th edition in 2012, and is also a proponent 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, it was just a random pattern that gave rise to an overfitted trading rule. We will illustrate this point with an example.

EXAMPLE. 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 interval [1, …, 22] for *Entry_day*, the interval [1, …, 20] for *Holding_period*, the interval [0, …, 10] for *Stop_loss*, and [-1, 1] for *Sign*. The 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.

[FIGURE 8 HERE]

Given the elevated Sharpe ratio, we could conclude that this strategy's performance is significantly greater than zero for any confidence level. Indeed, the *PSR-Stat* is 2.83, which implies a less than 1% probability that the true Sharpe ratio is below 0 (see Bailey and López de Prado [2012] for details). Several studies in the practitioners and 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. [2013].

In this practical application we have illustrated how simple is to produce overfit backtests when answering common investment questions, such as the presence of seasonal effects. We refer the reader to Appendix 4 for the implementation of this experiment in 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. In the context of regression overfitting, the great Enrico Fermi once remarked (Mayer et al. [2010]):

> *"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 bitter truth 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*. 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.

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.

APPENDICES

A.1. PROOF OF PROPOSITION 1

Embrechts et al. [2003, pp. 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, $z_n \sim \mathrm{Z}$, $n = 1, \ldots, N$, 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

$$\lim_{N\to\infty} Prob\left[\frac{max_N - \alpha}{\beta} \leq x\right] = \mathrm{G}[x] \tag{14}$$

where

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

The normalizing constants (α, β) are derived in Resnick [1987] and Embrechts et al. [2003]. 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 [1987]) is

$$\lim_{N\to\infty} E\left[\frac{max_N - \alpha}{\beta}\right] = \gamma \tag{15}$$

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

$$E[max_N] \approx \alpha + \gamma\beta = (1-\gamma)\mathrm{Z}^{-1}\left[1 - \frac{1}{N}\right] + \gamma\mathrm{Z}^{-1}\left[1 - \frac{1}{N}e^{-1}\right] \tag{16}$$

where *N>1*.■

A.2. PROOF OF PROPOSITION 2

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 $\mu_{IS}^{A} - \mu_{IS}^{B} = \underbrace{\frac{1-\delta}{\delta}}_{>0}(\mu_{OOS}^{B} - \mu_{OOS}^{A})$. 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

$$SR_{IS}^{A} > SR_{IS}^{B} \Leftrightarrow SR_{OOS}^{A} < SR_{OOS}^{B} \tag{17}$$

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). ■

A.3. PROOF OF PROPOSITION 3

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

$$m_{\tau} = (1-\varphi)\mu + \varphi m_{\tau-1} + \sigma\varepsilon_{\tau} \tag{18}$$

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 [2013] we know that the expected value of this process at a particular observation τ is

$$E_0[m_{\tau}] = (1-\varphi)\mu\frac{\varphi^{\tau} - 1}{\varphi - 1} + \varphi^{\tau}m_0 \tag{19}$$

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

$$E_0[m_{\tau}] = \frac{\mu + m_0}{2} \text{ ?} \tag{20}$$

Inserting Eq. (20) into Eq. (19), and solving for τ, we obtain

$$\tau = -\frac{Ln[2]}{Ln[\varphi]} \tag{21}$$

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

A.4. PROOF TO PROPOSITION 4

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 $\tau = 1, \dots, \delta T$, and the second subsample is called OOS, as it is comprised of $\tau = \delta T + 1, \dots, T$, 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 3, Eq. (18) we obtain

$$E_{\delta T}[m_T] - m_{\delta T} = \underbrace{(1 - \varphi^T)}_{>0}(\mu - m_{\delta T}) \tag{22}$$

because $\sigma_{IS}^{A} = \sigma_{IS}^{B}$, $SR_{IS}^{A} > SR_{IS}^{B} \Leftrightarrow m_{\delta T}^{A} > m_{\delta T}^{B}$. 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

$$SR_{IS}^{A} > SR_{IS}^{B} \Leftrightarrow SR_{OOS}^{A} < SR_{OOS}^{B} \tag{23}$$

.■

A.4. REPRODUCING THE RESULTS IN SECTION 9

The following Python code implements the experiment described in Section 9. The function *getRefDates_MonthBusinessDate* generates a list of dates for each business day of the month. These are the days that will be used for the *Entry_date* dimension of the four-dimensional mesh. Function *numBusinessDays* returns the number of business days between two dates. Function *getTrades* finds the trades associated with each parameter combination. *evalPerf* and *computePSR* evaluate the performance of the trading rule, while *attachTimeSeries* aligns the performance series for each of the 8800 nodes in the mesh. Finally *backTest* is the main function that carries out the experiment.

```
#!/usr/bin/env python
# On 20130704 by lopezdeprado@lbl.gov
import numpy as np,scipy.stats as ss,pandas as pd,datetime as dt
from random import gauss
from itertools import product
from sys import argv
#---------------------------------------------------------------------------------------
def getRefDates_MonthBusinessDate(dates):
	refDates,pDay={},[]
	first=dt.date(year=dates[0].year,month=dates[0].month,day=1)
	m=dates[0].month
	d=numBusinessDays(first,dates[0])+1
	for i in dates:
		if m!=i.month:m,d=i.month,1
		pDay.append(d)
		d+=1
	for j in range(1,30):
		lst=[dates[i] for i in range(len(dates)) if pDay[i]==j]
		refDates[j]=lst
	return refDates
#---------------------------------------------------------------------------------------
def numBusinessDays(date0,date1):
	m,date0_=1,date0
	while True:
		date0_+=dt.timedelta(days=1)
		if date0_>=date1:break
		if date0_.isoweekday()<6:m+=1
	return m
#---------------------------------------------------------------------------------------
def getTrades(series,dates,refDates,exit,stopLoss,side):
	# Get trades
	trades,pnl,position_,j,num=[],0,0,0,None
	for i in range(1,len(dates)):
		# Find next roll and how many trading dates to it
		if dates[i]>=refDates[j]:
			if dates[i-1]<refDates[j]:num,pnl=0,0
			if j<len(refDates)-1:
				while dates[i]>refDates[j]:j+=1
		if num==None:continue
		# Trading rule
		position=0
		if num<exit and pnl>stopLoss:position=side
		if position!=0 or position_!=0:
			trades.append([dates[i],num,position,position_*(series[i]-series[i-1])])
			pnl+=trades[-1][3]
			position_=position
		num+=1
	return trades
#---------------------------------------------------------------------------------------
def computePSR(stats,obs,sr_ref=0,moments=4):
	#Compute PSR
	stats_=[0,0,0,3]
	stats_[:moments]=stats[:moments]
	sr=stats_[0]/stats_[1]
	psrStat=(sr-sr_ref)*(obs-1)**0.5/(1-sr*stats_[2]+sr**2*(stats_[3]-1)/4.)**0.5
```

```
    psr=ss.norm.cdf((sr-sr_ref)*(obs-1)**0.5/(1-sr*stats_[2]+sr**2*(stats_[3]-1)/4.)**0.5)
    return psrStat,psr
#-----------------------------------------------------------------------------------------
def attachTimeSeries(series,series_,index=None,label='',how='outer'):
    # Attach a time series to a pandas dataframe
    if not isinstance(series_,pd.DataFrame):
        series_=pd.DataFrame({label:series_},index=index)
    elif label!='':series_.columns=[label]
    if isinstance(series,pd.DataFrame):
        series=series.join(series_,how=how)
    else:
        series=series_.copy(deep=True)
    return series
#-----------------------------------------------------------------------------------------
def evalPerf(pnl,date0,date1,sr_ref=0):
    freq=float(len(pnl))/((date1-date0).days+1)*365.25
    m1=np.mean(pnl)
    m2=np.std(pnl)
    m3=ss.skew(pnl)
    m4=ss.kurtosis(pnl,fisher=False)
    sr=m1/m2*freq**.5
    psr=computePSR([m1,m2,m3,m4],len(pnl),sr_ref=sr_ref/freq**.5,moments=4)[0]
    return sr,psr,freq
#-----------------------------------------------------------------------------------------
def backTest(nDays,factor):
    #1) Input parameters --- to be changed by the user
    holdingPeriod,sigma,stopLoss,length=20,1,10,1000
    #2) Prepare series
    date_,dates=dt.date(year=2000,month=1,day=1),[]
    while len(dates)<length:
        if date_.isoweekday()<5:dates.append(date_)
        date_+=dt.timedelta(days=1)
    series=np.empty((length))
    for i in range(series.shape[0]):
        series[i]=gauss(0,sigma)
        pDay_=dt.date(year=dates[i].year,month=dates[i].month,day=1)
        if numBusinessDays(pDay_,dates[i])<=nDays:
            series[i]+=sigma*factor
    series=np.cumsum(series)
    #3) Optimize
    refDates=getRefDates_MonthBusinessDate(dates)
    psr,sr,trades,sl,freq,pDay,pnl,count=None,None,None,None,None,None,None,0
    for pDay_ in refDates.keys():
        refDates_=refDates[pDay_]
        if len(refDates_)==0:continue
        #4) Get trades
        for prod_ in product(range(holdingPeriod+1),range(-stopLoss,1),[-1,1]):
            count+=1
            trades_=getTrades(series,dates,refDates_,prod_[0],prod_[1]*sigma, \
                prod_[2])
            dates_,pnl_=[j[0] for j in trades_],[j[3] for j in trades_]
            #5) Eval performance
            if len(pnl_)>2:
                #6) Reconcile PnL
                pnl=attachTimeSeries(pnl,pnl_,dates_,count)
                #7) Evaluate
```

```
                                    sr_,psr_,freq_=evalPerf(pnl_,dates[0],dates[-1])
                                    for j in range(1,len(pnl_)):pnl_[j]+=pnl_[j-1]
                                    if sr==None or sr_>sr:
                                            psr,sr,trades=psr_,sr_,trades_
                                            freq,pDay,prod=freq_,pDay_,prod_
                                            print count,pDay,prod,round(sr,2), \
                                                    round(freq,2),round(psr,2)
        print 'Total # iterations='+str(count)
        return pnl,psr,sr,trades,freq,pDay,prod,dates,series
#------------------------------------------------------------
# Boilerplate
if __name__=='__main__': backTest()
```

Snippet 1 – Evaluation of trading rules that search for Seasonal Effects

FIGURES

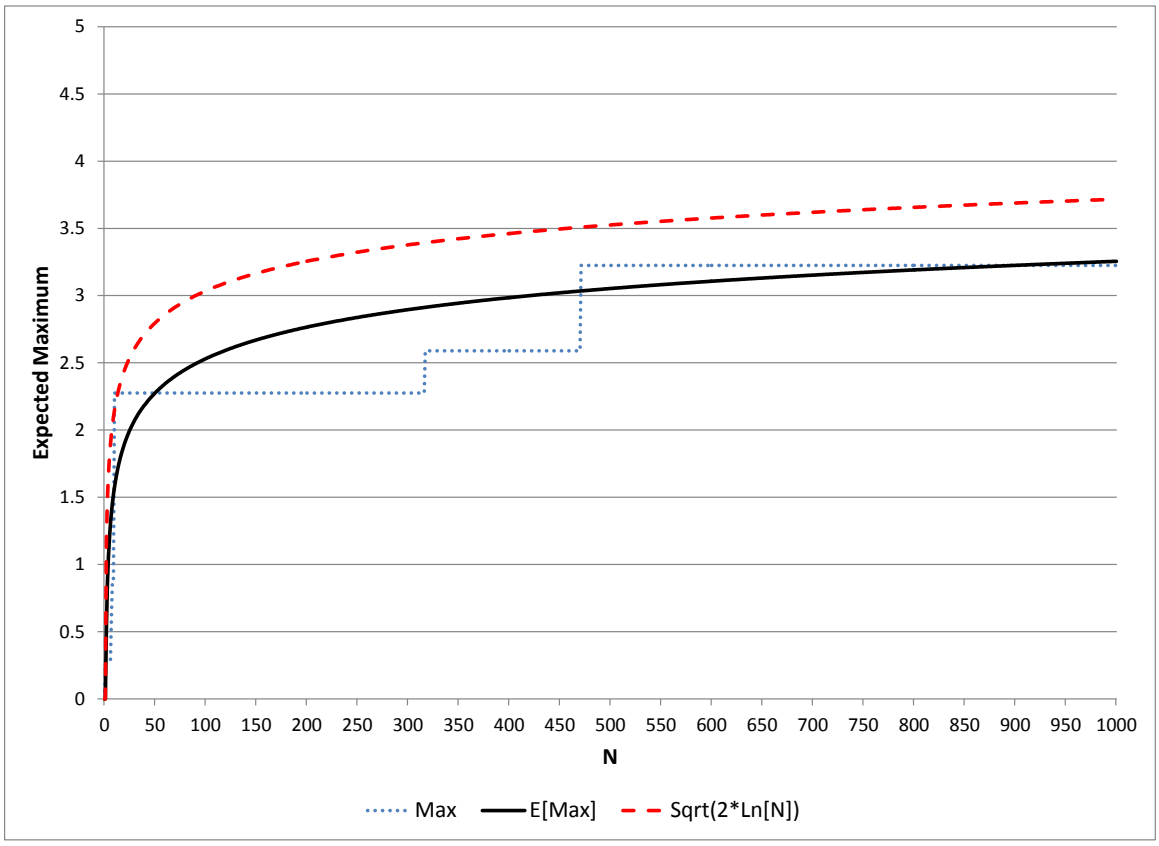

Figure 1 – Overfitting a backtest results as the number of trials grows

Figure 1 provides a graphical representation of Proposition 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.

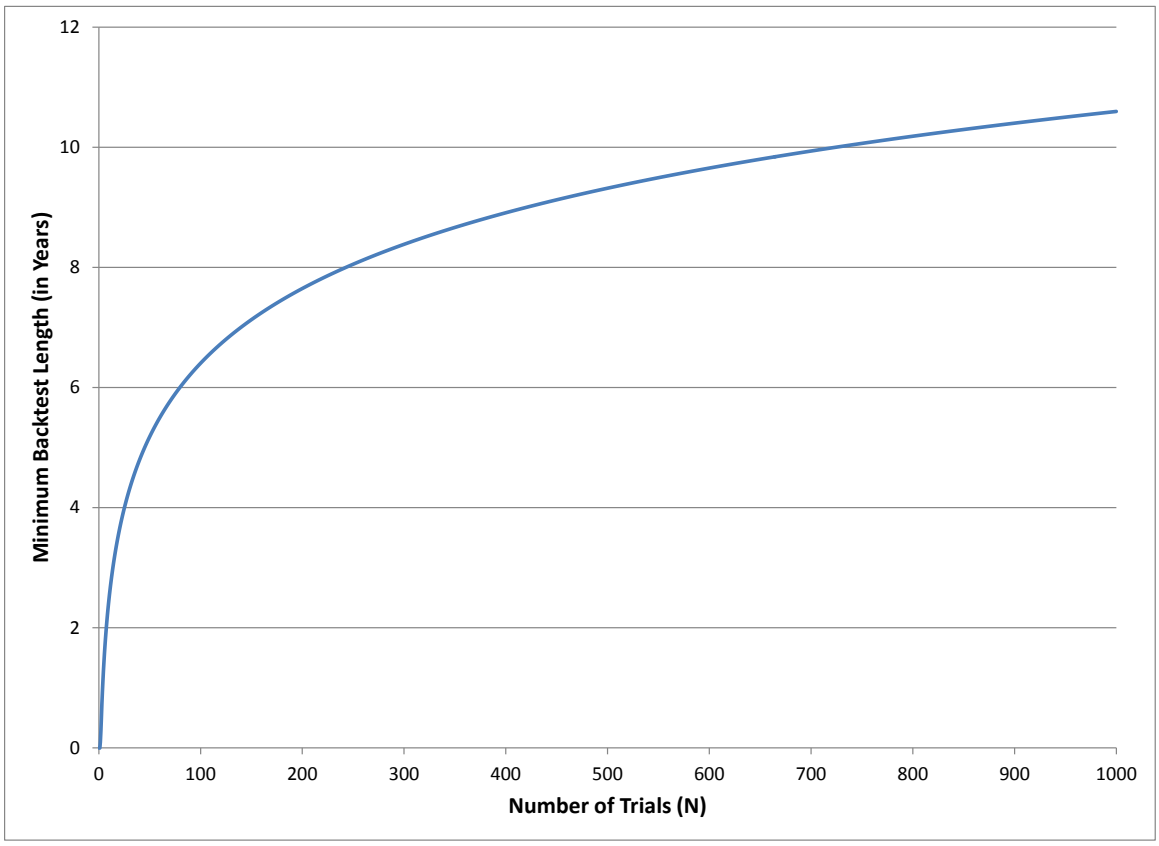

Figure 2 – Minimum Backtest Length needed to avoid overfitting, as a function of the number of trials

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.

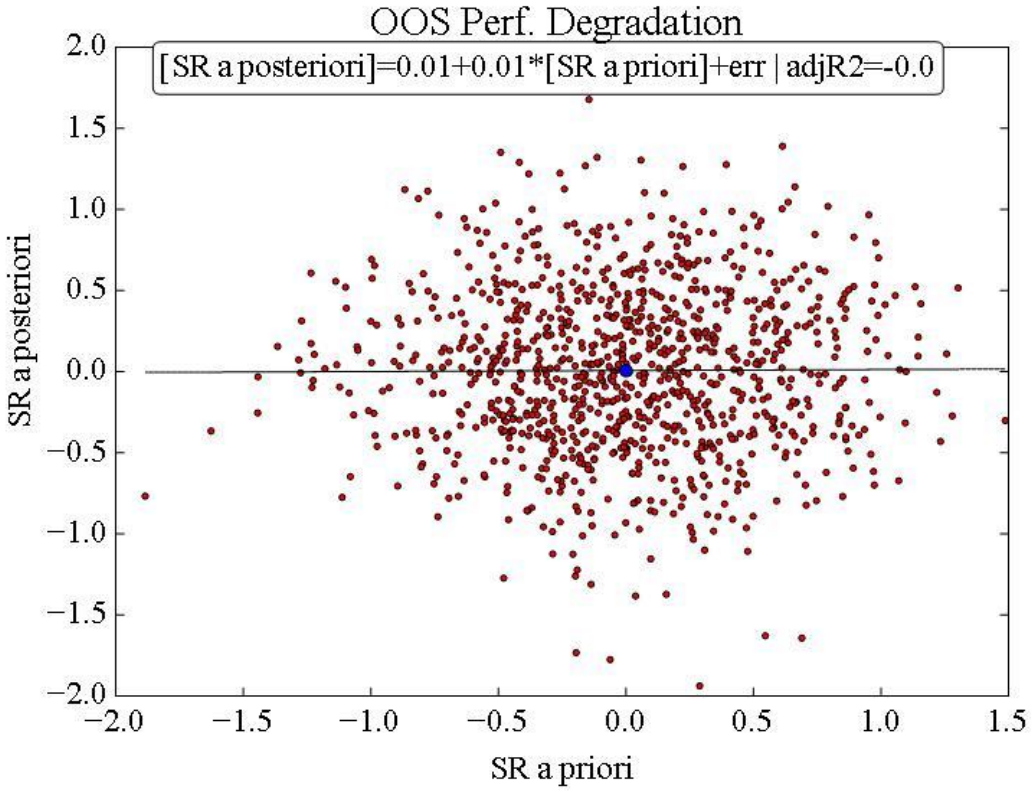

Figure 3 – Performance IS vs. OOS before introducing strategy selection

Figure 3 shows the relation between SR IS (x-axis) and SR OOS (y-axis), for μ=0,=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.

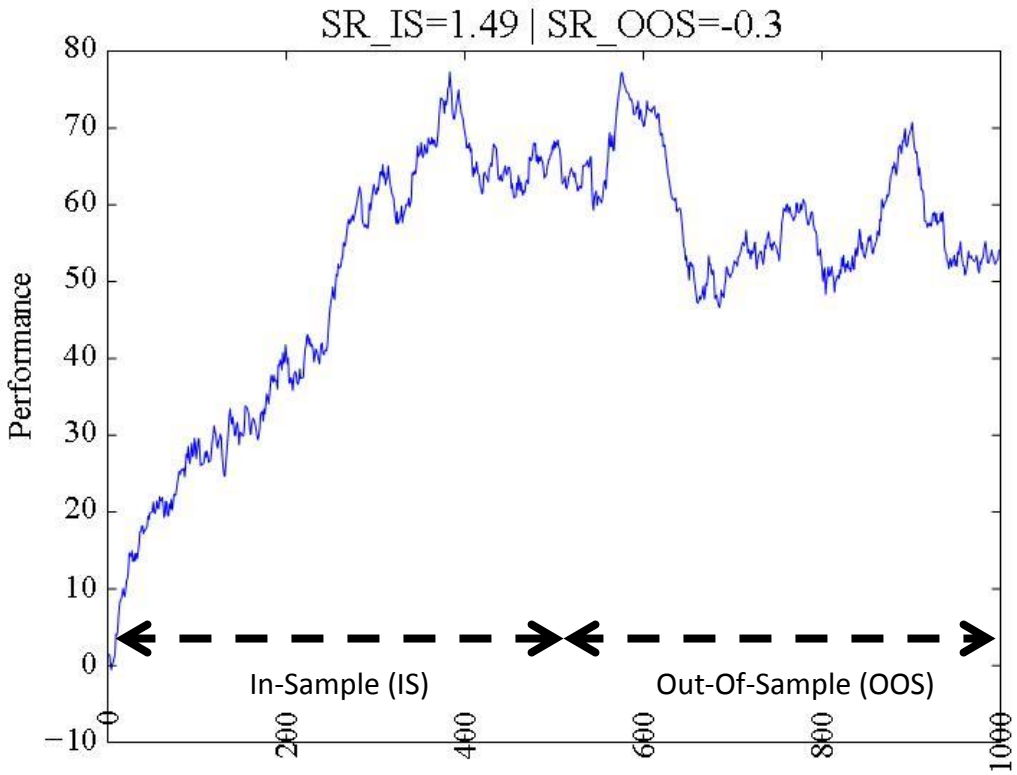

Figure 4 – Performance IS vs. performance OOS for one path, after introducing strategy selection

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.

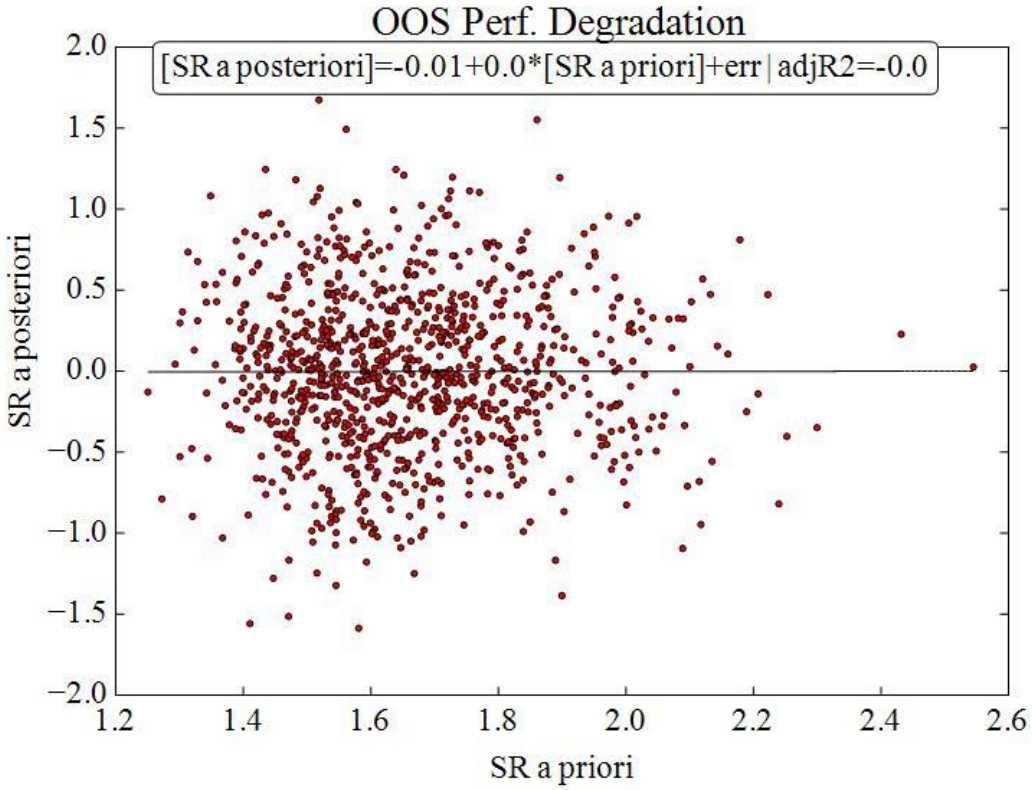

Figure 5 – Performance degradation after introducing strategy selection, in absence of compensation effects

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.

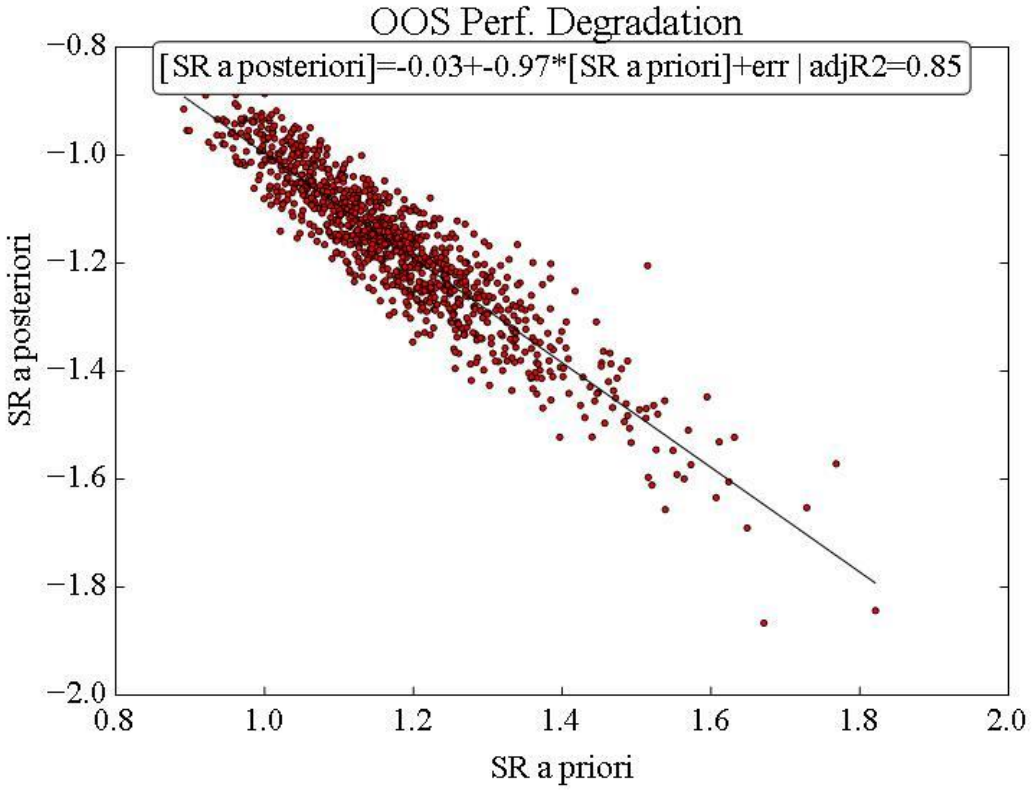

Figure 6 – Performance degradation as a result of strategy selection under compensation effects (global constraint)

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.

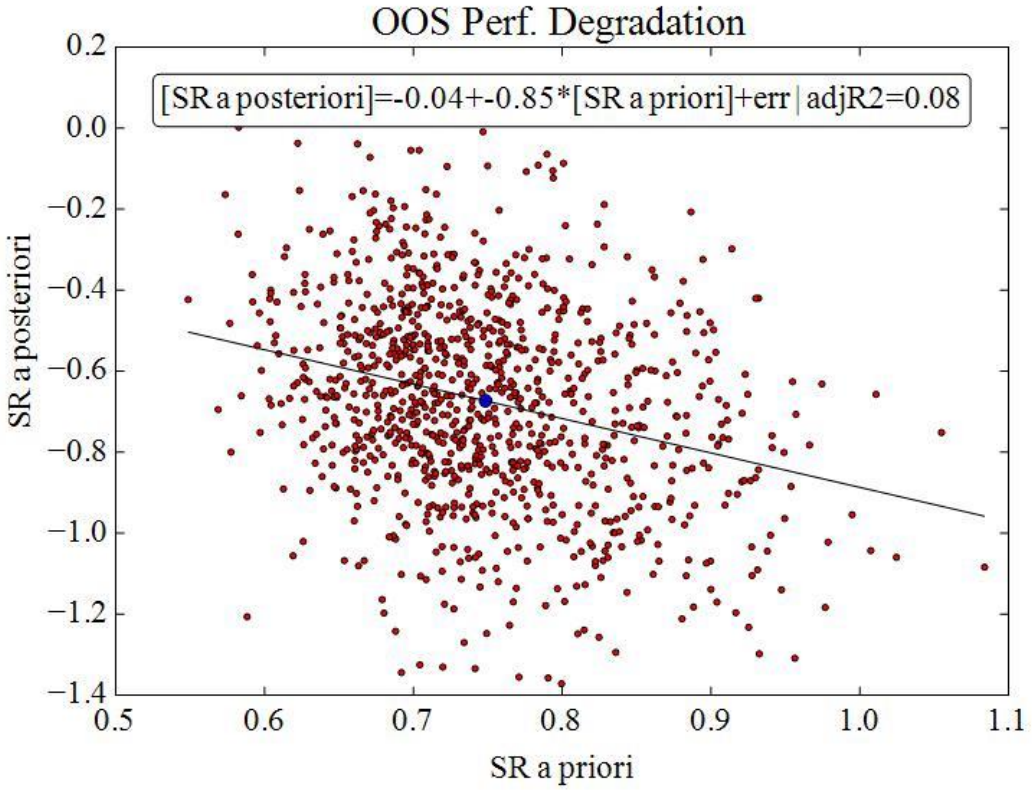

Figure 7 – Performance degradation as a result of strategy selection under compensation effects (first-order serial correlation)

Serially-correlated performance introduces another form of compensation effects, just as we saw in the case of a global constraint. For example, if φ=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.

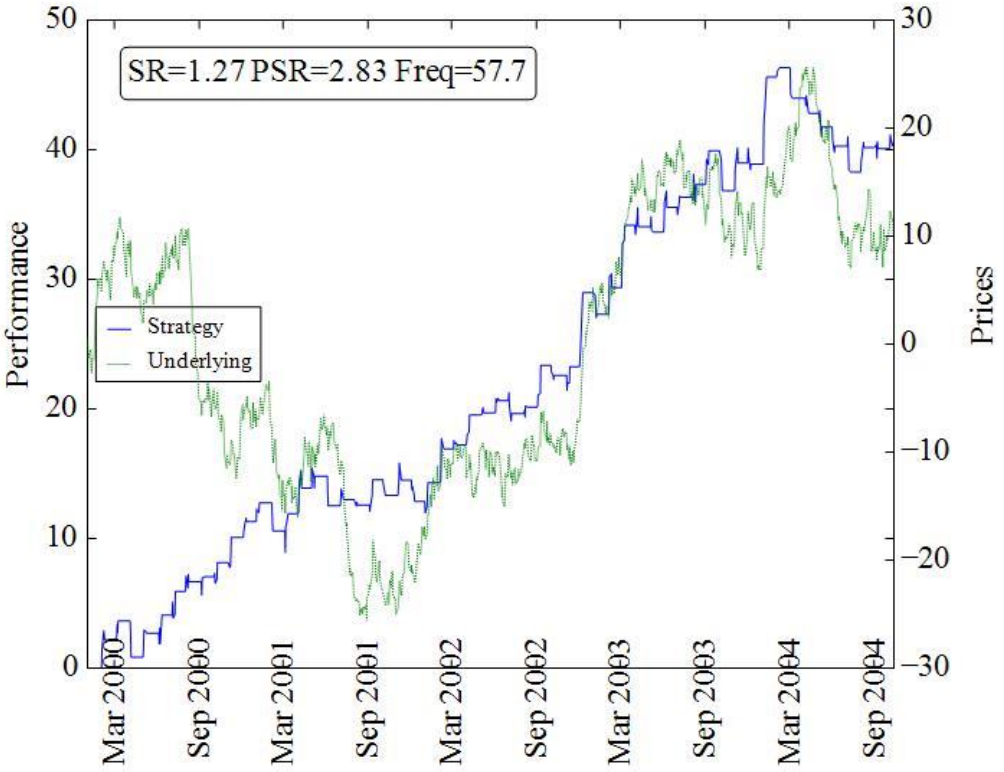

Figure 8 – Backtested performance of a seasonal strategy (example 1)

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.

REFERENCES

- Bailey, D., J. Borwein, M. López de Prado and J. Zhu (2013): "Computing the Probability of Backtest Overfitting." Working paper. Available at http://ssrn.com/abstract=2326253.
- Bailey, D. and M. López de Prado (2012): "The Sharpe Ratio Efficient Frontier," *Journal of Risk*, 15(2), pp. 3-44. Available at http://ssrn.com/abstract=1821643.
- Doyle, J. and C. Chen (2009): "The wandering weekday effect in major stock markets," *Journal of Banking & Finance*, 33, pp. 1388-1399.
- Embrechts, P., C. Klueppelberg and T. Mikosch (2003): "Modelling Extremal Events," Springer Verlag, New York.
- Feynman, R. (1964): "The Character of Physical Law," The MIT Press.
- Hadar, J. and W. Russell (1969): "Rules for Ordering Uncertain Prospects," *American Economic Review*, Vol. 59, pp. 25-34.
- Harvey, C., Y. Liu and H. Zhu (2013): "…and the Cross-Section of Expected Returns," Working Paper. SSRN.
- Harris, L. (2003): "Trading & Exchanges: Market Microstructure for Practitioners," Oxford University Press.
- Hawkins, D. (2004): "The problem of overfitting," *Journal of Chemical Information and Computer Science*, Vol. 44, pp. 1-12.
- Hirsch, Y. (1987): "Don't Sell Stocks on Monday," Penguin Books, 1st Edition.
- Leinweber, D. and K. Sisk (2011): "Event Driven Trading and the 'New News'," *Journal of Portfolio Management*, Vol. 38(1), 110-124.
- Lo, A. (2002): "The Statistics of Sharpe Ratios," *Financial Analysts Journal*, (58)4, July/August.
- López de Prado, M. and A. Peijan (2004): "Measuring the Loss Potential of Hedge Fund Strategies," *Journal of Alternative Investments*, Vol. 7(1), pp. 7-31. Available at http://ssrn.com/abstract=641702.
- López de Prado, M. and M. Foreman (2012): "A Mixture of Gaussians approach to Mathematical Portfolio Oversight: The EF3M algorithm," working paper, RCC at Harvard University. Available at http://ssrn.com/abstract=1931734.
- Mayer, J., K. Khairy and J. Howard (2010): "Drawing an Elephant with Four Complex Parameters," *American Journal of Physics*, 78(6).
- Resnick, S. (1987): "Extreme Values, Regular Variation and Point Processes," Springer.
- Schorfheide, F. and K. Wolpin (2012): "On the Use of Holdout Samples for Model Selection," *American Economic Review*, 102(3), pp. 477-481.
- Stodden, V., Bailey, D., Borwein, J., LeVeque, R, Rider, W. and Stein, W. (2013): "Setting the default to reproducible: Reproduciblity in computational and experimental mathematics," February, 2013. Available at http://www.davidhbailey.com/dhbpapers/icerm-report.pdf.
- Van Belle, G. and K. Kerr (2012): "Design and Analysis of Experiments in the Health Sciences," John Wiley & Sons.

- Weiss, S. and C. Kulikowski (1990): "Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems," Morgan Kaufman, 1st Edition.