THE DEFLATED SHARPE RATIO: CORRECTING FOR SELECTION BIAS, BACKTEST OVERFITTING AND NON-NORMALITY

David H. Bailey [†]

Marcos López de Prado [‡]

First version: April 15, 2014
This version: July 31, 2014

Journal of Portfolio Management, Forthcoming, 2014

[†] Recently retired from Lawrence Berkeley National Laboratory, Berkeley, CA 94720. Research Fellow at the University of California, Davis, Department of Computer Science. E-mail: david@davidhbailey.com

[‡] Senior Managing Director, Guggenheim Partners, New York, NY 10017. Research Affiliate, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. E-mail: lopezdeprado@lbl.gov

Special thanks are owed to Prof. David J. Hand (Royal Statistical Society), who reviewed an early version of this paper and suggested various extensions. We are also grateful to Matthew Beddall (Winton Capital), José Blanco (Credit Suisse), Jonathan M. Borwein (University of Newcastle, Australia), Sid Browne (Credit Suisse), Peter Carr (Morgan Stanley, NYU), Marco Dion (J.P. Morgan), Matthew D. Foreman (University of California, Irvine), Stephanie Ger (Berkeley Lab), Campbell Harvey (Duke University), Kate Land (Winton Capital), Jeffrey S. Lange (Guggenheim Partners), Attilio Meucci (KKR, NYU), Philip Protter (Columbia University), Riccardo Rebonato (PIMCO, University of Oxford), Mark Roulston (Winton Capital), Luis Viceira (HBS), John Wu (Berkeley Lab) and Jim Qiji Zhu (Western Michigan University).

THE DEFLATED SHARPE RATIO: CORRECTING FOR SELECTION BIAS, BACKTEST OVERFITTING AND NON-NORMALITY

ABSTRACT

With the advent in recent years of large financial data sets, machine learning and high-performance computing, analysts can backtest millions (if not billions) of alternative investment strategies. Backtest optimizers search for combinations of parameters that maximize the simulated historical performance of a strategy, leading to *backtest overfitting*.

The problem of performance inflation extends beyond backtesting. More generally, researchers and investment managers tend to report only positive outcomes, a phenomenon known as *selection bias*. Not controlling for the number of trials involved in a particular discovery leads to over-optimistic performance expectations.

The *Deflated Sharpe Ratio* (DSR) corrects for two leading sources of performance inflation: Selection bias under multiple testing and non-Normally distributed returns. In doing so, DSR helps separate legitimate empirical findings from statistical flukes.

Keywords: Sharpe ratio, Non-Normality, Probabilistic Sharpe ratio, Backtest overfitting, Minimum Track Record Length, Minimum Backtest Length.

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

Today's quantitative teams routinely scan through petabytes of financial data, looking for patterns invisible to the naked eye. In this endeavor, they are assisted by a host of technologies and advancing mathematical fields. Big data, machine learning, cloud networks and parallel processing have meant that millions (if not billions) of analyses can be carried out on a given dataset, searching for profitable investment strategies. To put this in perspective, the amount of data used by most quant teams today is comparable to the memory stored by Netflix to support its video-streaming business nationwide. This constitutes a radical change compared to the situation a couple of decades ago, when the typical financial analyst would run elementary arithmetic calculations on a spreadsheet containing a few thousand datapoints. In this paper we will discuss some of the unintended consequences of utilizing scientific techniques and high-performance computing without controlling for selection bias. While these problems are not specific to finance, examples are particularly abundant in financial research.

Backtests are a case in point. A backtest is a historical simulation of how a particular investment strategy would have performed in past. Although backtesting is a powerful and necessary research tool, it can also be easily manipulated. In this article we will argue that the most important piece of information missing from virtually all backtests published in academic journals and investment offerings is the number of trials attempted. Without this information, it is impossible to assess the relevance of a backtest. Put bluntly, *a backtest where the researcher has not controlled for the extent of the search involved in his or her finding is worthless, regardless of how excellent the reported performance might be.* Investors and journal referees should demand this information whenever a backtest is submitted to them, although even this will not remove the danger completely.

MULTIPLE TESTING

Investment strategies are typically judged according to performance statistics. Because any measurement is associated with a margin of error, the process of selecting a statistical model is an instance of decision-making under uncertainty. We can never be certain that the true performance is above a certain threshold, even if the estimated performance is. Two sorts of errors arise: The *Type I Error*, with probability α (also called "significance level"), and the *Type II Error*, with probability β. The Type I Error occurs when we choose a strategy that should have been discarded (a "false positive"), and the Type II Error occurs when we discard a strategy that should have been chosen (a "false negative"). Decision makers are often more concerned with "false positives" than with "false negatives". The reason is, they would rather exclude a true strategy than risking the addition of a false one. For risk-averse investors, a lost opportunity is less worrisome than an actual loss. For this reason, the standard practice is to design statistical tests which set the Type I Error probability to a low threshold (e.g., $\alpha = 5\%$), while maximizing the power of the test, defined as $1 - \beta$.

Suppose now that we are interested in analyzing multiple strategies on the same dataset, with the aim of choosing the best, or at least a good one, for future application. A curious problem then emerges: As we test more and more strategies, each at the same significance level α, the overall probability of choosing at least one poor strategy grows. This is called the multiple testing problem, and it is so pervasive and notorious that the American Statistical Society explicitly warns against it its Ethical Guidelines (American Statistical Association, 1997, guideline #8):

Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading.

SELECTION BIAS

Researchers conducting multiple tests on the same data tend to publish only those that pass a statistical significance test, hiding the rest. Because negative outcomes are not reported, investors are only exposed to a biased sample of outcomes. This problem is called "selection bias", and it is caused by multiple testing combined with partial reporting, see Roulston and Hand [2013]. It appears in many different forms: Analysts who do not report the full extent of the experiments conducted ("file drawer effect"), journals that only publish "positive" outcomes ("publication bias"), indices that only track the performance of hedge funds that didn't blow up ("survivorship bias"), managers who only publish the history of their (so far) profitable strategies ("self-selection bias", "backfilling"), etc. What all these phenomena have in common is that critical information is hidden from the decision-maker, with the effect of a much larger than anticipated Type I Error probability. Ignoring the full extent of trials makes the improbable more probable, see Hand [2014].

The danger of encountering false positives is evident in High-Throughput Screening (HTS) research projects, such as those utilized to discover drug treatments, the design of chemical compounds or microarrays genetic testing. Bennett et al. [2010] were awarded the 2012 Ig Nobel prize for showing that even a salmon's dead brain can appear to show significant activity under multiple MRI testing. More seriously, the pharmaceutical field has been stung with numerous recent instances of products that look great based on published trials, but which disappoint when actually fielded. The problem here, in many cases, is that only the results of successful tests are typically published, thereby introducing a fundamental bias into the system. Such experiences have led to the AllTrials movement (see http://alltrials.net), which would require the results of all trials to be made public.

In spite of the experience of HTS projects, it is rare to find financial studies that take into account the increased false positive rates that result from hiding multiple testing. The extent of this problem has led some researchers to paraphrase Ioannidis [2005] and conclude that "most claimed research findings in financial economics are likely false", see Harvey et al. [2013].

BACKTEST OVERFITTING

What constitutes a legitimate empirical finding? After a sufficient number of trials, it is guaranteed that a researcher will always find a misleadingly profitable strategy, a false positive. Random samples contain patterns, and a systematic search through a large space of strategies will eventually lead to identifying one that profits from the chance configuration of how the random data have fallen. When a set of parameters are optimized to maximize the performance of a backtest, an investment strategy is likely to be fit to such flukes. This phenomenon is called

"backtest overfitting", and we refer the reader to Bailey et al. [2014] for a detailed discussion. Although the historical performance of an optimized backtest may seem promising, the random pattern that fuels it is unlikely to repeat itself in the future, hence rendering the strategy worthless. Structural breaks have nothing to do with the failure of the strategy in this situation.

Let us elucidate this point with an example. After tossing a fair coin ten times we could obtain by chance a sequence such as {+,+,+,+,+,-,-,-,-,-}, where "+" means head and "-" means tail. A researcher could determine that the best strategy for betting on the outcomes of this coin is to expect "+" on the first five tosses, and for "-" on the last five tosses (a typical "seasonal" argument in the investment community). When we toss that coin ten more times, we may obtain a sequence such as {-,-,+,-,+,+,-,-,+,-}, where we win 5 times and lose 5 times. That researcher's betting rule was overfit, because it was designed to profit from a random pattern present only in the past. The rule has null power over the future, regardless of how well it appears to have worked in the past.

Competition among investment managers means that the ratio of signal to noise in financial series is low, increasing the probability of "discovering" a chance configuration, rather than an actual signal. The implication is that backtest overfitting is hard to avoid. Clearly this is a critical issue, because most investment decisions involve choosing among multiple candidates or alternatives.

AN ONLINE TOOL TO EXPLORE BACKTEST OVERFITTING

Researchers at the Lawrence Berkeley National Laboratory have developed an online application to explore the phenomenon of backtest overfitting. It first generates a pseudorandom time series mimicking a history of stock market prices. It then finds the parameter combination (holding period, stop loss, entry day, side, etc.) that optimizes the strategy's performance. The tool usually has no problem finding a "profitable" strategy with any desired Sharpe ratio. Yet when this "profitable" strategy is applied to a second similar-length pseudorandom time series, it typically flounders, producing little gain or even a loss. To try this tool, visit http://datagrid.lbl.gov/backtest.

BACKTEST OVERFITTING UNDER MEMORY EFFECTS

In order to understand the effects of backtest overfitting on out-of-sample performance, we must introduce an important distinction: Financial processes with and without memory. A coin, whether fair or biased, does not have memory. The 50% heads ratio does not arise as a result of the coin "remembering" previous tosses. Patterns emerge, but they are "diluted away" as additional sequences of tosses are produced. Now suppose that we add a memory chip to that coin, such that it remembers the previous tosses and distributes its mass to compensate its outcomes. This memory actively "undoes" recent historical patterns, such that the 50% heads ratio is quickly recovered. Just as a spring memorizes its equilibrium position, financial variables that have acquired a high tension will return to equilibrium violently, undoing previous patterns.

The difference between "diluting" and "undoing" a historical pattern is enormous. Diluting does not contradict your bet, but undoing generates outcomes that systematically go against your bet!

Backtest overfitting tends to identify the trading rules that would profit from the most extreme random patterns in sample. In presence of memory effects, those extreme patters must be undone, which means that backtest overfitting will lead to loss maximization. See Bailey et al. (2014) for a formal mathematical proof of this statement. Unfortunately, most financial series exhibit memory effects, a situation that makes backtest overfitting a particularly onerous practice, and may explain why so many systematic funds fail to perform as advertised.

BACKTEST OVERFITTING AND THE HOLDOUT METHOD

Practitioners attempt to validate their backtests using several approaches. The holdout method is perhaps the best known example, see Schorfheide and Wolpin [2012] for a description. A researcher splits the available sample into two non-overlapping subsets: The in-sample subset (IS), and the out-of-sample subset (OOS). The idea is to discover a model using the IS subset, and then validate its generality on the OOS subset. The k-fold cross-validation method repeats the process of sample splitting *k* times, which contributes to the reduction of the estimation error's variance. Then, OOS results are tested for statistical significance. For example, we could reject models where the OOS performance is inconsistent with the IS performance.

From our earlier discussion, the reader should understand why the holdout method can*not* prevent backtest overfitting: Holdout assesses the generality of a model as if a single trial had taken place, again ignoring the rise in false positives as more trials occur. If we apply the holdout method enough times (say 20 times for a 95% confidence level), false positives are no longer unlikely: They are expected. The more times we apply holdout, the more likely an invalid strategy will pass the test, which will then be published as a single-trial outcome. While model validation techniques are relevant for guarding against testing hypotheses suggested by the data (called "Type III errors"), they do not control for backtest overfitting.

GENERAL APPROACHES TO MULTIPLE TESTING

In the previous sections we have introduced the problem of multiple testing. We have explained how multiple testing leads to an increased probability of false positives, and hiding negative outcomes of multiple testing leads to selection bias. We have seen that backtest overfitting is a particularly expensive form of selection bias, as a result of memory effects present in financial series. Finally, we have discussed why popular model validation techniques fail to address these problems. So what would constitute a proper strategy selection method?

Statisticians have been aware of the multiple testing and selection bias problems since the early part of the twentieth century, and have developed methods to tackle them (e.g. the classic Bonferroni approach to multiple testing, and Heckman's work on selection bias - which won him a Nobel Prize). Recently, however, with the increase in large data sets and, in particular, challenges arising from bioinformatics, tackling multiple testing problems has become a hot research topic, with many new advances being made. See, for example, Dudoit and van der Laan [2008] and Dmitrienko et al. [2010]. For controlling *familywise error rate* - the probability that at least one test of multiple null hypotheses will be falsely rejected - a variety of methods have been developed, but researchers have also explored alternative definitions of error rate (e.g., Holm [1979]). In particular, the complementary *false discovery rate* (e.g. Benjamini and

Hochberg [1995]) has attracted a great deal of attention. Instead of looking at the probability of falsely rejecting a true null hypothesis, this looks at the probability that a rejected hypothesis is null (which is arguably more relevant in most scientific situations, if not in the manufacturing situations for which Neyman and Pearson originally developed their approach to hypothesis testing).

Bailey et al. [2013] introduce a new cross-validation technique to compute the *Probability of Backtest Overfitting* (PBO). PBO assesses whether the strategy selection process has been conducive to overfitting, in the sense that selected strategies tend to underperform the median of trials out of sample. PBO is non-parametric and can be applied to any performance statistic, however it requires a large amount of information. We dedicate the remainder of the paper to develop a new parametric method to correct the Sharpe ratio for the effects of multiple testing, inspired by the false discovery rate approach.

EXPECTED SHARPE RATIOS UNDER MULTIPLE TRIALS

The Sharpe Ratio (SR) is the most widely used performance statistic (Sharpe [1966, 1975, 1994]). It evaluates an investment in terms of returns on risk, as opposed to return on capital. Portfolio managers are keen to improve their SRs, in order to rank higher in databases such as Hedge Fund Research, and receive greater capital allocations. Setting a constant cut-off threshold for SR above which portfolio managers or strategies are considered for capital allocation leads to the same selection bias discussed earlier: As more candidates are considered, the false positive rate keeps growing.

More formally, consider a set of N independent backtests or track records associated with a particular strategy class (e.g., Discretionary Macro). Each element of the set is called a *trial*, and it is associated with a SR estimate, $\widehat{SR}_n$, with $n = 1, \dots, N$. Suppose that these trials' $\{\widehat{SR}_n\}$ follow a Normal distribution, with mean $E[\{\widehat{SR}_n\}]$ and variance $V[\{\widehat{SR}_n\}]$. This is not an unreasonable assumption, since the concept of "strategy class" implies that the trials are bound by some common characteristic pattern. In other words, we assume that there is a mean and variance associated with the trials' $\{\widehat{SR}_n\}$ for a given strategy class. For example, we would expect the $E[\{\widehat{SR}_n\}]$ from High Frequency Trading trials to be greater than $E[\{\widehat{SR}_n\}]$ from Discretionary Macro. Appendix 1 proves that, under these assumptions, the expected maximum of $\{\widehat{SR}_n\}$ after $N \gg 1$ independent trials can be approximated as:

$$E[max\{\widehat{SR}_n\}] \approx E[\{\widehat{SR}_n\}] + \sqrt{V[\{\widehat{SR}_n\}]}\left((1-\gamma)Z^{-1}\left[1-\frac{1}{N}\right] + \gamma Z^{-1}\left[1-\frac{1}{N}e^{-1}\right]\right) \quad (1)$$

where γ (approx. 0.5772) is the Euler-Mascheroni constant, Z is the cumulative function of the standard Normal distribution, and e is Euler's number. Appendix 2 runs numerical experiments that contrast the accuracy of this solution. Appendix 3 shows how N can be determined when the trials are not independent.

Equation 1 tells us that, as the number of independent trials (N) grows, so will grow the expected maximum of $\{\widehat{SR}_n\}$. Exhibit 1 illustrates this point for $E[\{\widehat{SR}_n\}] = 0$, $V[\{\widehat{SR}_n\}] = 1$ and $N \in [10,1000]$.

[EXHIBIT 1 HERE]

Consequently, it is not surprising to obtain good backtest results or meet better portfolio managers as we parse through more candidates. This is a consequence of purely random behavior, because we will observe better candidates *even if there is no investment skill associated with this strategy class* ($E[\{\widehat{SR}_n\}] = 0, V[\{\widehat{SR}_n\}] > 0$). In the following section we will use this fact to adjust the strategy rejection threshold as the number of independent trials increases.

THE DEFLATED SHARPE RATIO

When an investor selects the best performing strategy over a large number of alternatives, she exposes herself to "the winner's curse." As we have shown in the previous section, she is likely to choose a strategy with an inflated Sharpe ratio. Performance out of sample is likely to disappoint, a phenomenon called "regression to the mean" in the shrinkage estimation literature, see Efron [2011]. In what follows we will provide an estimator of the Sharpe ratio that undoes the selection bias introduced by multiple testing, while also correcting for the effects of Non-Normal returns.

The *Probabilistic Sharpe Ratio* (PSR), developed in Bailey and López de Prado [2012a], computes the probability that the true SR is above a given threshold. This rejection threshold is determined by the user. PSR takes into account the sample length and the first four moments of the returns' distribution. The reason for this, as several studies have demonstrated, is the inflationary effect of short samples and samples drawn from non-Normal returns distributions. We refer the interested reader to Lo [2002], Mertens [2002], López de Prado and Peijan [2004], Ingersoll et al. [2007] for a discussion.

Our earlier analysis has shown a second source of inflation, caused by selection bias. Both sources of inflation, when conflated, can lead to extremely high estimated values $\widehat{SR}$, even when the true SR may be null. In this paper we propose a *Deflated Sharpe Ratio* (DSR) statistic that corrects for both sources of $\widehat{SR}$ inflation, defined as:

$$\widehat{DSR} \equiv \widehat{PSR}(\widehat{SR}_0) = Z\left[\frac{(\widehat{SR} - \widehat{SR}_0)\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}\right] \tag{2}$$

where $\widehat{SR}_0 = \sqrt{V[\{\widehat{SR}_n\}]}\left((1-\gamma)Z^{-1}\left[1 - \frac{1}{N}\right] + \gamma Z^{-1}\left[1 - \frac{1}{N}e^{-1}\right]\right)$, $V[\{\widehat{SR}_n\}]$ is the variance across the trials' estimated SR and N is the number of independent trials. We also use information concerning the selected strategy: $\widehat{SR}$ is its estimated SR, T is the sample length, $\hat{\gamma}_3$ is

the skewness of the returns distribution and $\hat{\gamma}_4$ is the kurtosis of the returns distribution for the selected strategy. *Z* is the cumulative function of the standard Normal distribution.

Essentially, DSR is a PSR where the rejection threshold is adjusted to reflect the multiplicity of trials. The rationale behind DSR is the following: Given a set of SR estimates, $\{\widehat{SR}_n\}$, its expected maximum is greater than zero, even if the true SR is zero. Under the null hypothesis that the actual Sharpe ratio is zero, $H_0: SR = 0$, we know that the expected maximum $\widehat{SR}$ can be estimated as the $\widehat{SR}_0$ in Eq.(2). Indeed, $\widehat{SR}_0$ increases as more independent trials are attempted (*N*), or the trials involve a greater variance $\left(V\left[\{\widehat{SR}_n\}\right]\right)$.

Note that the standard $\widehat{SR}$ is computed as a function of two estimates: Mean and standard deviation of returns. DSR deflates SR by taking into consideration five additional variables: The non-Normality of the returns $(\hat{\gamma}_3, \hat{\gamma}_4)$, the length of the returns series (T), the variance of the SRs tested $\left(V\left[\{\widehat{SR}_n\}\right]\right)$, as well as the number of independent trials involved in the selection of the investment strategy (N).

In an excellent recent study, Harvey and Liu [2014], henceforth denoted HL, compute the threshold that a new strategy's Sharpe ratio must overcome in order to evidence greater performance. HL's solution is based on Benjamini and Hochberg's framework. The role of HL's threshold is analogous to the role played by our $E\left[max\{\widehat{SR}_n\}\right]$ in Eq. (1), which we derived through Extreme Value Theory. DSR uses this threshold to deflate a particular Sharpe ratio estimate (see Eq. (2)). In other words, DSR computes how statistically significant a particular $\widehat{SR}$ is, considering the set of trials carried out so far. In the current paper we apply DSR to the $E\left[max\{\widehat{SR}_n\}\right]$ threshold, but DSR could also be computed on HL's threshold. From that perspective, these two methods are complementary, and we encourage the reader to compute DSR using both thresholds, $E\left[max\{\widehat{SR}_n\}\right]$ as well as HL's.

A NUMERICAL EXAMPLE

Suppose that a strategist is researching seasonality patterns in the treasury market. He believes that the U.S. Treasury's auction cycle creates inefficiencies that can be exploited by selling off-the-run bonds a few days before the auction, and buying the new issue a few days after the auction. He backtests alternative configurations of this idea, by combining different pre-auction and post-auction periods, tenors, holding periods, stop-losses, etc. He uncovers that many combinations yield an annualized $\widehat{SR}$ of 2, with a particular one yielding a $\widehat{SR}$ of 2.5 over a daily sample of 5 years.

Excited by this result, he calls an investor asking for funds to run this strategy, arguing that an annualized $\widehat{SR}$ of 2.5 must be statistically significant. The investor, who is familiar with a paper recently published by the *Journal of Portfolio Management*, asks the strategist to disclose: i) The number of independent trials carried out (*N*); ii) the variance of the backtest results $\left(V\left[\{\widehat{SR}_n\}\right]\right)$; iii) the sample length (*T*); and iv) the skewness and kurtosis of the returns $(\hat{\gamma}_3, \hat{\gamma}_4)$. The analyst responds that $N = 100$, $V\left[\{\widehat{SR}_n\}\right] = \frac{1}{2}$, *T=1250*, $\hat{\gamma}_3 = -3$ and $\hat{\gamma}_4 = 10$.

Shortly after, the investor declines the analyst's proposal. Why? Because the investor has determined that this is not a legitimate empirical discovery at a 95% confidence level. In particular, $\widehat{SR}_0 = \sqrt{\frac{1}{2 \cdot 250}}\left((1-\gamma)Z^{-1}\left[1-\frac{1}{100}\right]+\gamma Z^{-1}\left[1-\frac{1}{100}e^{-1}\right]\right) \approx 0.1132$, non-annualized (with 250 observations per year), and $\widehat{DSR} \approx Z\left[\frac{\left(\frac{2.5}{\sqrt{250}}-0.1132\right)\sqrt{1249}}{\sqrt{1-(-3)\frac{2.5}{\sqrt{250}}+\frac{10-1}{4}\left(\frac{2.5}{\sqrt{250}}\right)^2}}\right] = 0.9004 <$ 0.95.

[EXHIBIT 2 HERE]

Exhibit 2 plots how the rejection threshold $\widehat{SR}_0$ increases with *N*, and consequently $\widehat{DSR}$ decreases. The investor has recognized that there is only a 90% chance that the true SR associated with this strategy is greater than zero. Should the strategist have made his discovery after running only *N=46* independent trials, the investor may have allocated some funds, as $\widehat{DSR}$ would have been 0.9505, above the 95% confidence level.

Non-Normality also played a role in discarding this investment offer. If the strategy had exhibited Normal returns ($\hat{\gamma}_3 = 0, \hat{\gamma}_4 = 3$), $\widehat{DSR} = 0.9505$ after *N=88* independent trials. If non-Normal returns had not inflated the performance so much, the investor would have been willing to accept a much larger number of trials. This example illustrates that it is critical for investors to account for both sources of performance inflation jointly, as DSR does.

WHEN SHOULD WE STOP TESTING?

One important practical implication of this research is that multiple testing is a useful tool that should not be abused. Multiple testing exercises should be carefully planned in advance, so as to avoid running an unnecessarily large number of trials. Investment theory, not computational power, should motivate what experiments are worth conducting. This begs the question, what is the optimal number of trials that should be attempted?

An elegant answer to this critical question can be found in the theory of optimal stopping, more concretely the so called "secretary problem", or 1/e-law of optimal choice, see Bruss [1984]. There are many versions of this problem, but the key notion is that we wish to impose a cost to the number of trials conducted, because every additional trial irremediably increases the probability of a false positive.[1]

In the context of our discussion, it translates as follows: From the set of strategy configurations that are theoretically justifiable, sample a fraction $1/e$ of them (roughly 37%) at random and measure their performance. After that, keep drawing and measuring the performance of additional configurations from that set, one by one, until you find one that beats all of the

[1] In the original "secretary problem" the decision-maker had no possibility of choosing a past candidate. This is not necessarily the case when selecting a strategy. However the key similarity is that in both cases the counter of trials cannot be turned back.

previous. That is the optimal number of trials, and that "best so far" strategy the one that should be selected.

This result provides a useful rule of thumb, with applications that go beyond the number of strategy configurations that should be backtested. It can be applied to situations where we test multiple alternatives with the goal of choosing a near-best *as soon as possible*, so as to minimize the chances of a false positive.

CONCLUSIONS

Machine learning, high-performance computing and related technologies have advanced many fields of the sciences. For example, the U.S. Department of Energy's SciDAC (Scientific Discovery through Advanced Computing) program uses terascale computers to *"research problems that are insoluble by traditional theoretical and experimental approaches, hazardous to study in the laboratory, or time-consuming or expensive to solve by traditional means."*[2]

Many of these techniques have become available to financial analysts, who use them to search for profitable investment strategies. Academic journals often publish backtests that report the performance of such simulated strategies. One problem with these innovations is that, unless strict scientific protocols are followed, there is a substantial risk of selecting and publishing false positives.

We have argued that selection bias is ubiquitous in the financial literature, where backtests are often published without reporting the full extent of the trials involved in selecting that particular strategy. To make matters worse, we know that backtest overfitting in the presence of memory effects leads to negative performance out-of-sample. Thus, selection bias combined with backtest overfitting misleads investors into allocating capital to strategies that will systematically lose money. The customary disclaimer that "past performance does not guarantee future results" is too lenient when in fact adverse outcomes are very likely.

In this paper we have proposed a test to determine whether an estimated SR is statistically significant after correcting for two leading sources of performance inflation: Selection bias and non-Normal returns. The Deflated Sharpe Ratio (DSR) incorporates information about the unselected trials, such as the number of independent experiments conducted and the variance of the SRs, as well as taking into account the sample length, skewness and kurtosis of the returns' distribution.

[2] http://outreach.scidac.gov/scidac-overview

APPENDICES

A.1. DERIVING THE EXPECTED MAXIMUM SHARE RATIO

We would like to derive the expected Sharpe Ratio after *N* independent trials. Consider a set $\{y_n\}$ of independent and identically distributed random variables drawn from a Normal Distribution, $y_n \sim \mathcal{N}[\mu, \sigma^2]$, $n = 1, \ldots, N$. We can build a standardized set $\{x_n\}$ by computing $x_n \equiv \frac{y_n - \mu}{\sigma}$, where $x_n \sim \mathcal{N}[0,1] \equiv Z$. The set $\{y_n\}$ is therefore identical to the set $\{\mu + \sigma x_n\}$. For $\sigma > 0$, the order of the set $\{\mu + \sigma x_n\}$ is unchanged, consequently

$$max\{y_n\} = max\{\mu + \sigma x_n\} = \mu + \sigma max\{x_n\} \tag{3}$$

where the same element is the maximum in both sets. Because the mathematical expectation operator, $E[.]$, is linear, we know that

$$E[max\{y_n\}] = \mu + \sigma E[max\{x_n\}] \tag{4}$$

Bailey et al. [2014a] prove that given a series of independent and identically distributed standard normal random variables, $x_n \sim Z$, $n = 1, \ldots, N$, the expected maximum of that series, $E[max_N] \equiv E[max\{x_n\}]$, can be approximated for a large *N* as

$$E[max_N] \approx (1-\gamma) Z^{-1}\left[1 - \frac{1}{N}\right] + \gamma Z^{-1}\left[1 - \frac{1}{N} e^{-1}\right] \tag{5}$$

where γ (approx. 0.5772) is the Euler-Mascheroni constant, e is Euler's number, and $N \gg 1$. Combining both results, we obtain that

$$E[max\{y_n\}] \approx \mu + \sigma\left((1-\gamma) Z^{-1}\left[1 - \frac{1}{N}\right] + \gamma Z^{-1}\left[1 - \frac{1}{N} e^{-1}\right]\right) \tag{6}$$

■.

A.2. EXPERIMENTAL VERIFICATION

We can evaluate experimentally the accuracy of the previous result. First, for a given set of parameters $\{E[\{\widehat{SR}_n\}], V[\{\widehat{SR}_n\}], N\}$ we compute $E[max\{\widehat{SR}_n\}]$ analytically, i.e. applying Eq. (6). Second, we would like to compare that value with a numerical estimation. That numerical estimation is obtained by drawing *Q* random sets of $\{\widehat{SR}_n\}$ of size *N* from a Normal distribution with mean $E[\{\widehat{SR}_n\}]$ and variance $V[\{\widehat{SR}_n\}]$, computing the maximum of those sets, and estimating their mean, $Q^{-1}\sum_{q=1}^{Q} max\left[\{\widehat{SR}_n\}_q\right]$. Third, we compute the estimation error as $\varepsilon = E[max\{\widehat{SR}_n\}] - Q^{-1}\sum_{q=1}^{Q} max\left[\{\widehat{SR}_n\}_q\right]$, where we expect $\varepsilon \approx 0$. Fourth, we can repeat the previous three steps for a wide variety of combinations of $\{E[\{\widehat{SR}_n\}], V[\{\widehat{SR}_n\}], N\}$, and evaluate the magnitude and patterns of the resulting ε under various scenarios.

[EXHIBIT 3.1 HERE]

[EXHIBIT 3.2 HERE]

Exhibit 3.1 plots a heat map of ε values where $E\left[\{\widehat{SR}_n\}\right] \in [-10,-10]$, $N \in [10,1000]$, $Q = 10^4$ and $V\left[\{\widehat{SR}_n\}\right] = 1$. Exhibit 3.2 plots the analogous heat map, where we have set $V\left[\{\widehat{SR}_n\}\right] = 4$. It is easy to verify that alternative values of $V\left[\{\widehat{SR}_n\}\right] > 0$ generate similar outcomes. Snippet 1 implements in Python the experiment discussed earlier.

```
#!/usr/bin/env python
# On 20140607 by lopezdeprado@lbl.gov
import numpy as np,scipy.stats as ss,pandas as pd
from itertools import product
#---------------------------------------------------------------------------
def getExpMaxSR(mu,sigma,numTrials):
   # Compute the expected maximum Sharpe ratio (Analytically)
   emc=0.5772156649 # Euler-Mascheroni constant
   maxZ=(1-emc)*ss.norm.ppf(1-1./numTrials)+emc*ss.norm.ppf(1-1./(numTrials*np.e))
   return mu+sigma*maxZ
#---------------------------------------------------------------------------
def getDistMaxSR(mu,sigma,numTrials,numIters):
   # Compute the expected maximum Sharpe ratio (Numerically)
   maxSR,count=[],0
   while count<numIters:
      count+=1
      series=np.random.normal(mu,sigma,numTrials)
      maxSR.append(max(series))
   return np.mean(maxSR),np.std(maxSR)
#---------------------------------------------------------------------------
def simulate(mu,sigma,numTrials,numIters):
   #1) Compute Expected[Max{Sharpe ratio}] ANALYTICALLY
   expMaxSR=getExpMaxSR(mu,sigma,numTrials)
   #2) Compute Expected[Max{Sharpe ratio}] NUMERICALLY
   meanMaxSR,stdMeanMaxSR=getDistMaxSR(mu,sigma,numTrials,numIters)
   return expMaxSR,meanMaxSR,stdMeanMaxSR
#---------------------------------------------------------------------------
def main():
   numIters,sigma,output,count=1e4,1,[],0
   for prod_ in product(np.linspace(-100,100,101),range(10,1001,10)):
      mu,numTrials=prod_[0],prod_[1]
      expMaxSR,meanMaxSR,stdMaxSR=simulate(mu,sigma,numTrials,numIters)
      err=expMaxSR-meanMaxSR
      output.append([mu,sigma,numTrials,numIters,expMaxSR,meanMaxSR, \
         stdMaxSR,err])
   output=pd.DataFrame(output,columns=['mu','sigma','numTrials','numIters', \
      'expMaxSR','meanMaxSR','stdMaxSR','err'])
   output.to_csv('DSR.csv')
   return
#---------------------------------------------------------------------------
if __name__=='__main__':
   main()
```

Snippet 1 – Code for the experimental verification

A.3. ESTIMATING THE NUMBER OF INDEPENDENT TRIALS

It is critical to understand that the N used to compute $E[max\{SR_n\}]$ corresponds to the number of *independent* trials. Suppose that we run M trials, where only N trials are independent, $N<M$. Clearly, using M instead of N will overstate $E[max\{SR_n\}]$. So given M dependent trials we need to derive the number of "implied independent trials", $\widehat{N}$.

One path to accomplish that is by taking into account the average correlation between the trials, ρ. First, consider an MxM correlation matrix C with real-valued entries $\{\rho_{i,j}\}$, where i is the index for the rows and j is the index for the columns. Let $\tilde{C}$ be a modified correlation matrix, where all off-diagonal correlations have been replaced by a constant value ρ, i.e. $\rho_{i,j} = \rho, \forall i \neq j$. Then we define the weighted average correlation as the value ρ such that

$$x'Cx = x' \begin{bmatrix} 1 & \rho_{1,j} & \rho_{1,M} \\ \rho_{i,1} & 1 & \rho_{i,M} \\ \rho_{M,1} & \rho_{M,j} & 1 \end{bmatrix} x = x' \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} x = x'\tilde{C}x \tag{7}$$

In words, we are interested in finding the constant value ρ such that, if we make all off-diagonal correlations equal to ρ, the above quadratic form remains unchanged. For the case where x equals a unit vector, $x = 1_M$, the quadratic form reduces to the sum of all the entries $\{\rho_{i,j}\}$, leading to the equal-weighted average correlation:

$$\rho = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M}\rho_{i,j} - M}{M(M-1)} = \frac{2\sum_{i=1}^{M}\sum_{j=i+1}^{M}\rho_{i,j}}{M(M-1)} \tag{8}$$

Second, a proper correlation matrix must be positive-definite, so it is guaranteed that all its quadratic forms are strictly positive, and in particular ${1_M}'C1_M = {1_M}'\tilde{C}1_M > 0$. Then, ${1_M}'\tilde{C}1_M = M + M(M-1)\rho$. The implication is that the average correlation is bounded by $\rho \in \left(-\frac{1}{M-1}, 1\right]$, with $M > 1$ for a correlation to exist. The larger the number of trials, the more positive the average correlation is likely to be, and for a sufficiently large M we have $-\frac{1}{M-1} \approx 0 < \rho \leq 1$.

Third, we know that as $\rho \to 1$, then $N \to 1$. Similarly, as $\rho \to 0$, then $N \to M$. Given an estimated average correlation $\hat{\rho}$, we could therefore interpolate between these two extreme outcomes to obtain

$$\widehat{N} = \hat{\rho} + (1 - \hat{\rho})M \tag{9}$$

Exhibit 4 plots the relationship between $\{M, \hat{\rho}, \widehat{N}\}$. This method could be further enriched by incorporating Fisher's transform (see Fisher [1915]), thus controlling for the variance of the error in the estimation of $\hat{\rho}$.

[EXHIBIT 4 HERE]

This and other "average correlation" approaches are intuitive and convenient in practical applications. However, two problematic aspects should be highlighted in connection with "average correlation" formulations: First, correlation is a limited notion of linear dependence. Second, in practice *M* almost always exceeds the sample length, *T*. Then the estimate of average correlation may itself be overfit. In general for short samples $T < \frac{1}{2}M(M-1)$, the correlation matrix will be numerically ill-conditioned, and it is not guaranteed that ${1_M}'C1_M > 0$. Estimating an average correlation is then pointless, because there are more correlations $\{\rho_{i,j} | i < j, i = 1, \ldots M\}$ than independent pairs of observations! One way to deal with this numerical problem is to reduce the dimension of the correlation matrix (see Bailey and López de Prado [2012b] for one such algorithm), and compute the average correlation on that reduced definite-positive matrix.

An alternative and more direct path is to use information theory to determine $\widehat{N}$. Entropy relates to a much deeper concept of redundancy than correlation. We refer the interested reader to the literature on *data compression*, *total correlation* and *multiinformation*, e.g. Watabane [1960] and Studený and Vejnarová [1999]. These and other standard information theory methods produce accurate estimates of the number *N* of non-redundant sources among a set of *M* random variables.

EXHIBITS

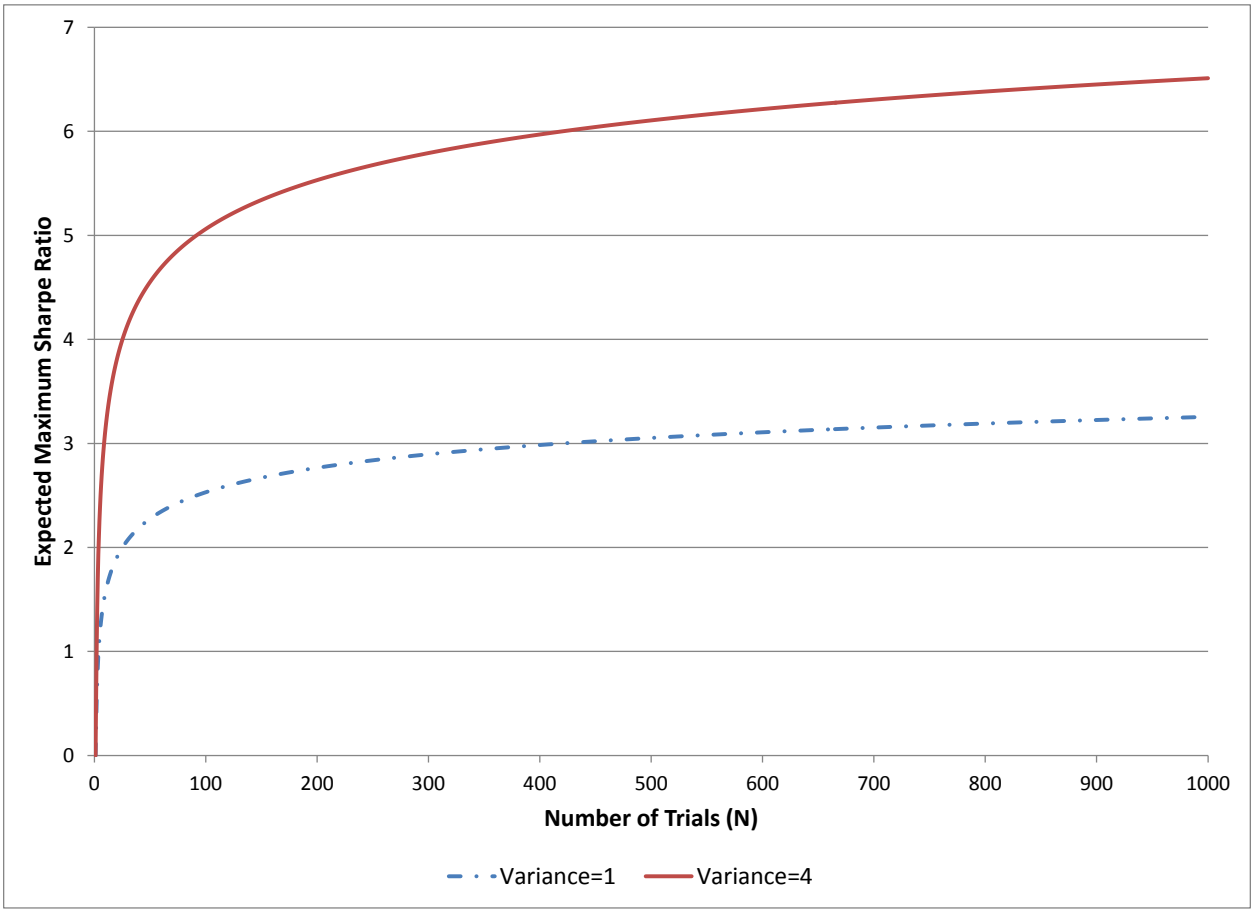

Exhibit 1 – Expected Maximum Sharpe Ratio as the number of independent trials grows, for $E\left[\left\{\widehat{SR}_n\right\}\right] = 0$ *and* $V\left[\left\{\widehat{SR}_n\right\}\right] \in \{1,4\}$

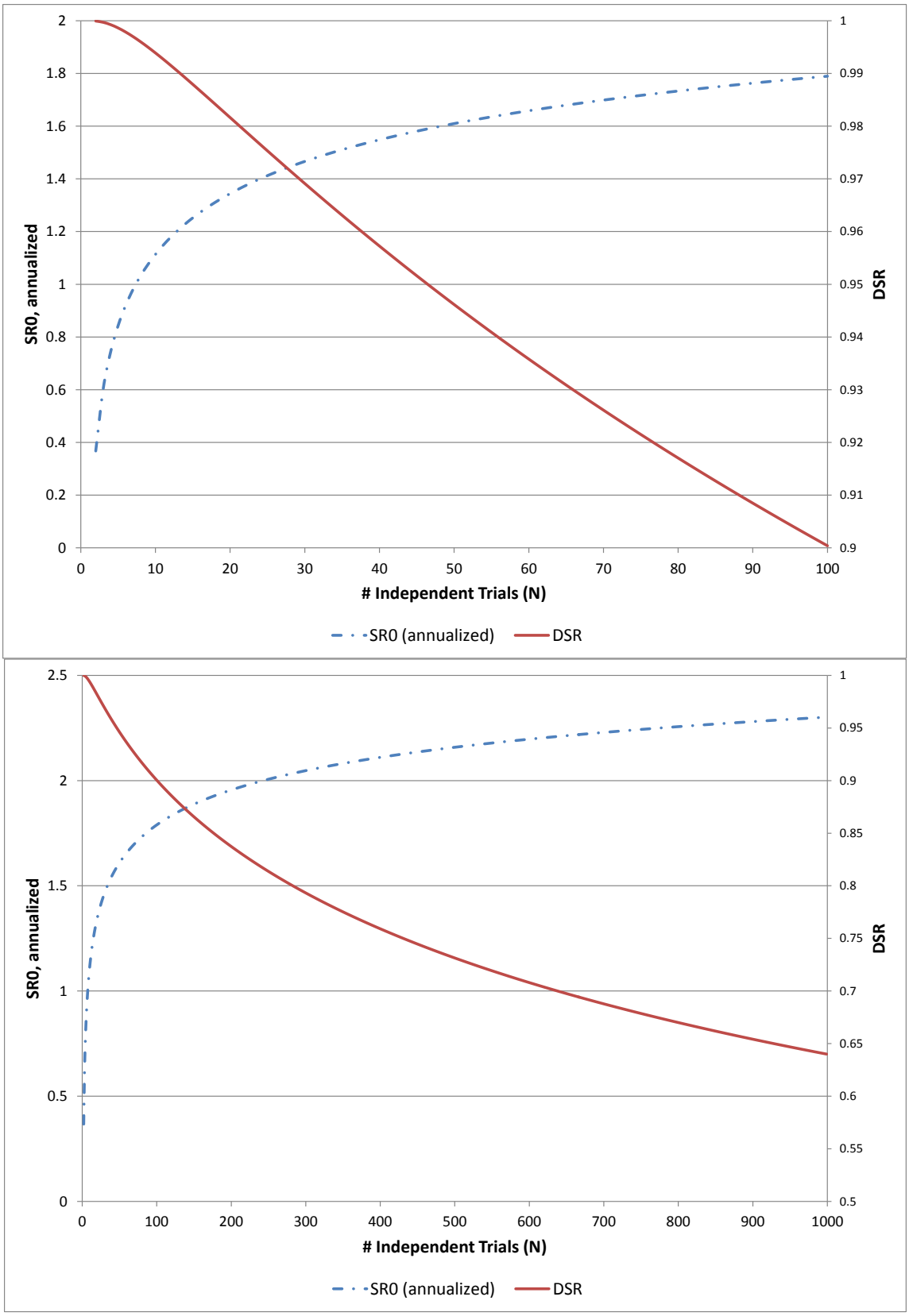

Exhibit 2 – Expected Maximum Sharpe Ratio (left y-axis) and Deflated Sharpe Ratio (right y-axis) as the number of independent trials grows

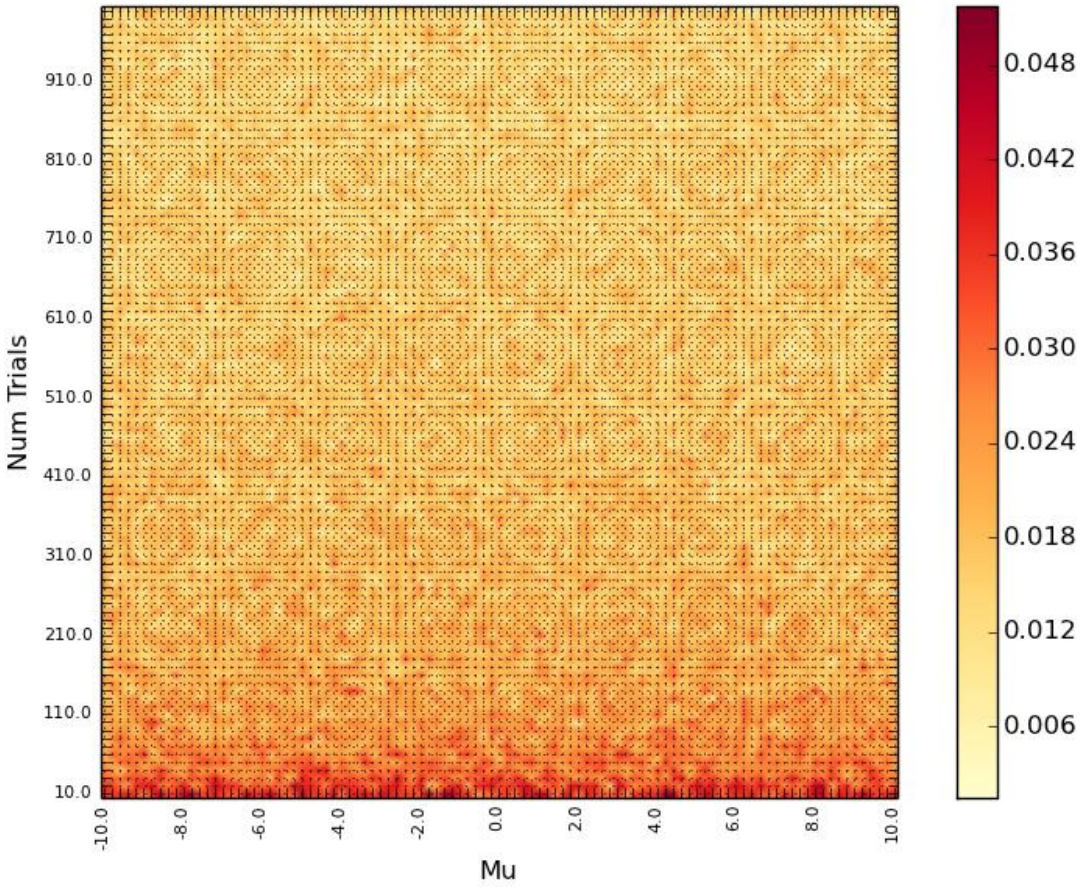

Exhibit 3.1 – Experimental verification of the analytical formula for estimating the expected maximum of $\{\widehat{SR}_n\}$*, where* $V[\{\widehat{SR}_n\}] = 1$

This heat map plots the difference between the expected maximum of $\{\widehat{SR}_n\}$ estimated analytically and the average of maximum $\{\widehat{SR}_n\}$ computed empirically, for various combinations of μ and N. We have set $V[\{\widehat{SR}_n\}] = 1$. When the number of trials is very small (e.g., less than 50), the analytical result sometimes overestimates the empirical result, by a very small margin (less than 0.05 for an underlying process with variance 1). As the number of trials increases, that level of overestimation converges to zero (by 1000 trials it is as small as 0.006 for a process with variance 1). This is consistent with our proof, where we required that $N \gg 1$.

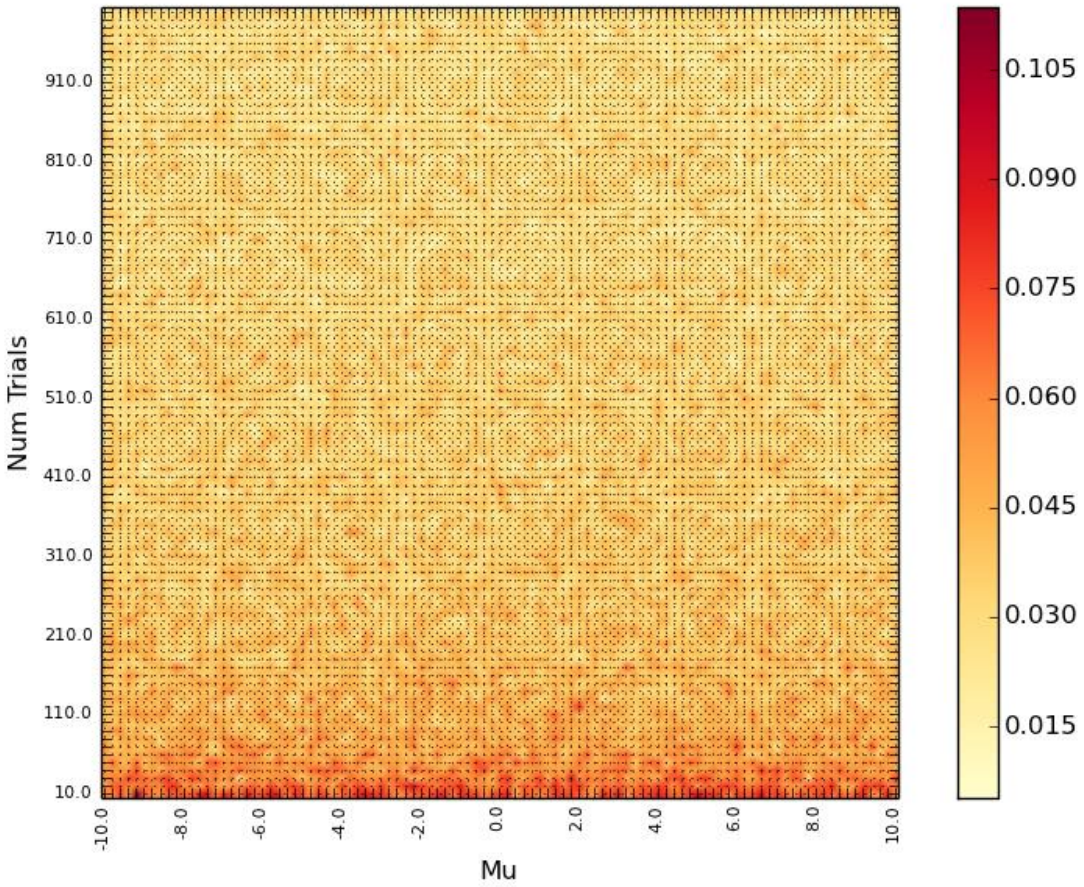

Exhibit 3.2 – Experimental verification of the analytical formula for estimating the expected maximum of $\{\widehat{SR}_n\}$*, where* $V[\{\widehat{SR}_n\}] = 4$

An analogous result is obtained for alternative values of $V[\{\widehat{SR}_n\}]$. Now the maximum error is approx. 0.11, for an underlying process with variance $V[\{\widehat{SR}_n\}] = 4$. This is about double the maximum error observed in Exhibit 2.1, and consistent with the $\sigma = \sqrt{V[\{\widehat{SR}_n\}]}$ scaling in Eq.(6). The error quickly converges to zero as the number of trials grows, $N \gg 1$.

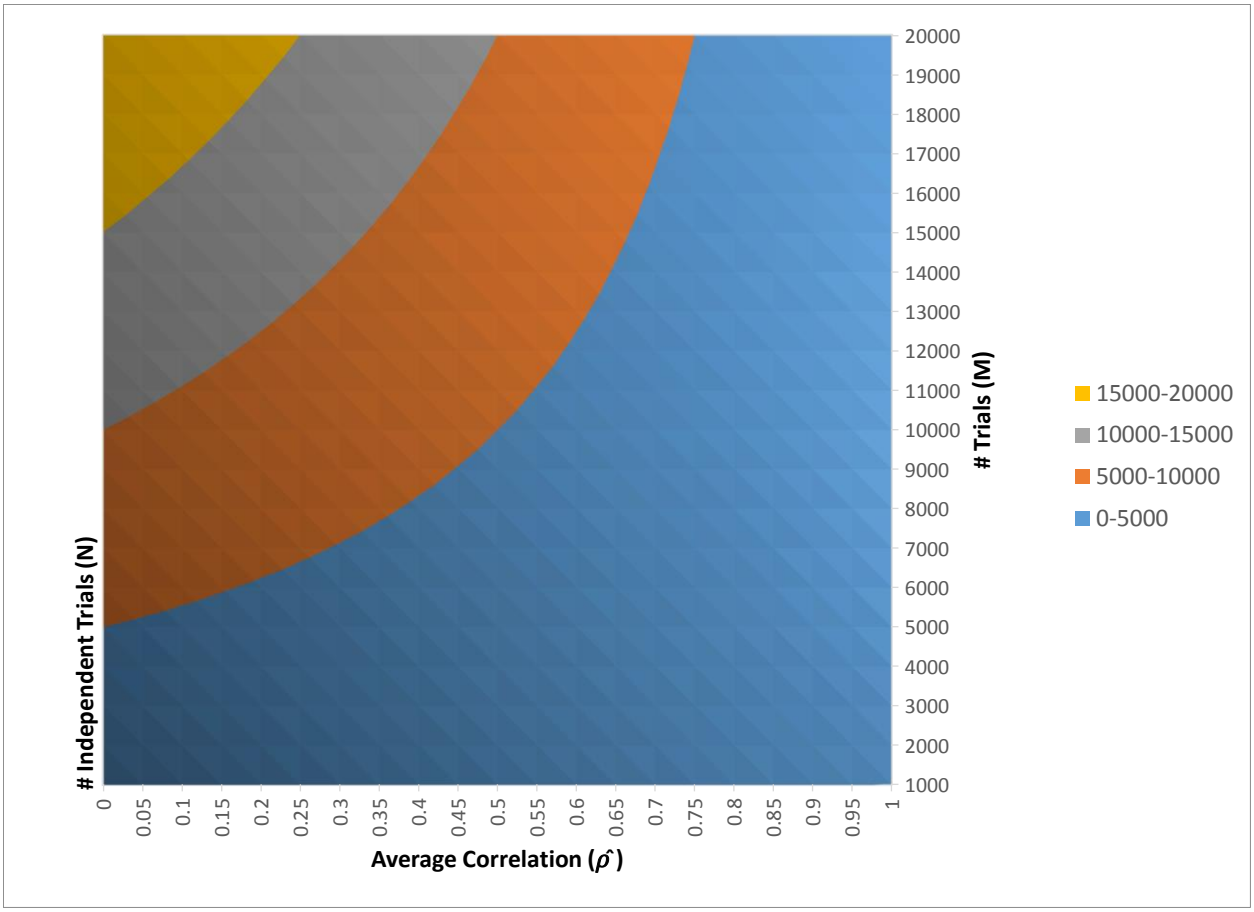

Exhibit 4 – Implied number of independent trials ($\hat{N}$) for various values of average correlation ($\hat{\rho}$) and number of trials (M)

REFERENCES

1. "Ethical Guidelines for Statistical Practice." American Statistical Society (1999). Available at: http://www.amstat.org/committees/ethics/index.html
2. Bailey, D., J. Borwein, M. López de Prado and J. Zhu. "The Probability of Backtest Overfitting." Working paper, SSRN (2013). Available at: http://ssrn.com/abstract=2326253
3. Bailey, D., J. Borwein, M. López de Prado and J. Zhu. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-Of-Sample Performance." Notices of the American Mathematical Society, Vol. 61, No. 5 (May 2014a). Available at: http://ssrn.com/abstract=2308659
4. Bailey, D. and M. López de Prado. "The Sharpe Ratio Efficient Frontier." Journal of Risk, Vol. 15, No. 2 (Winter, 2012a).
5. Bailey, D. and M. López de Prado. "Balanced Baskets: A New Approach to Trading and Hedging Risks." Journal of Investment Strategies (Risk Journals), Vol. 1, No. 4 (Fall, 2012b).
6. Beddall, M. and K. Land. "The hypothetical performance of CTAs". Working paper, Winton Capital Management (2013).
7. Benjamini, Y. and Y. Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." Journal of the Royal Statistical Society, Series B (Methodological), Vol. 57, No. 1 (1995), pp. 289 - 300.
8. Bennet, C., A. Baird, M. Miller and G. Wolford. "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction", Journal of Serendipitous and Unexpected Results, Vol. 1, No. 1 (2010), pp. 1-5.
9. Bruss, F. "A unified Approach to a Class of Best Choice problems with an Unknown Number of Options". Annals of Probability, Vol. 12, No. 3 (1984), pp. 882–891.
10. Dmitrienko, A., A.C. Tamhane, and F. Bretz (eds.) Multiple Testing Problems in Pharmaceutical Statistics. 1st edition, Boca Raton, FL: CRC Press, 2010.
11. Dudoit, S. and M.J. van der Laan. Multiple Testing Procedures with Applications to Genomics. 1st edition, Berlin: Springer, 2008.
12. Fisher, R.A. "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population." Biometrika (Biometrika Trust), Vol. 10, No. 4 (1915), pp. 507–521.
13. Hand, D. J. The Improbability Principle. 1st edition, New York, NY: Scientific American/Farrar, Straus and Giroux, 2014.
14. Harvey, C., Y. Liu and H. Zhu. "…And the Cross-Section of Expected Returns." Working paper, Duke University, 2013. Available at: http://ssrn.com/abstract=2249314
15. Harvey, C. and Y. Liu. "Backtesting." Working paper, Duke University, 2014. Available at http://ssrn.com/abstract=2345489
16. Hochberg Y. and A. Tamhane. Multiple comparison procedures. 1st edition, New York, NY: Wiley, 1987.
17. Holm, S. "A Simple sequentially rejective multiple test procedure." Scandinavian Journal of Statistics, Vol. 6 (1979), pp. 65-70.

18. Ioannidis, J.P.A. "Why most published research findings are false." PloS Medicine, Vol. 2, No. 8 (2005), pp. 696-701.
19. Ingersoll, J., M. Spiegel, W. Goetzmann, I. Welch. "Portfolio performance manipulation and manipulation-proof performance measures." The Review of Financial Studies, Vol. 20, No. 5 (2007), pp. 1504-1546.
20. Lo, A. "The Statistics of Sharpe Ratios." Financial Analysts Journal, Vol. 58, No. 4 (July/August 2002), pp. 36-52.
21. López de Prado, M., A. Peijan. "Measuring Loss Potential of Hedge Fund Strategies." Journal of Alternative Investments, Vol. 7, No. 1 (Summer, 2004), pp. 7-31. http://ssrn.com/abstract=641702
22. Mertens, E. "Variance of the IID estimator in Lo (2002)." Working paper, University of Basel, 2002.
23. Roulston, M. and D. Hand. "Blinded by Optimism." Working paper, Winton Capital Management, December 2013.
24. Schorfheide, F. and K. Wolpin. "On the Use of Holdout Samples for Model Selection." American Economic Review, Vol. 102, No. 3 (2012), pp. 477-481.
25. Sharpe, W. "Mutual Fund Performance." Journal of Business, Vol. 39, No. 1 (1966), pp. 119–138.
26. Sharpe, W. "Adjusting for Risk in Portfolio Performance Measurement." Journal of Portfolio Management, Vol. 1, No. 2 (Winter, 1975), pp. 29-34.
27. Sharpe, W. "The Sharpe ratio." Journal of Portfolio Management, Vol. 21, No. 1 (Fall, 1994), pp. 49-58.
28. Studený M. and Vejnarová J. "The multiinformation function as a tool for measuring stochastic dependence." in M I Jordan, ed., Learning in Graphical Models, Cambridge, MA: MIT Press, 1999, pp. 261–296,
29. Watanabe S. "Information theoretical analysis of multivariate correlation." IBM Journal of Research and Development, Vol. 4 (1960), pp. 66–82.