Limitations of Quantitative Claims About Trading Strategy Evaluation

Michael Harris
Price Action Lab
mikeh@priceactionlab.com

First draft: July 15, 2016

Abstract

One of the key assumptions of quantitative trading strategy evaluation is that Type II errors (missed discoveries) are preferable to Type I errors (false discoveries.) However, practitioners have known for long that the statistical properties of some genuine trading strategies are often indistinguishable from those of random trading strategies. Therefore, any adjustments of statistics to guard against p-hacking increase Type II error unless the power of the test is high. At the same time, the power of the test is limited by insufficient samples and changing market conditions. Furthermore, genuine strategies with statistical properties that are similar to those of random strategies may overfit due to favorable market conditions but fail when market conditions change. These facts severely limit the effectiveness of quantitative claims about trading strategy evaluation. Practitioners have instead resorted to Monte Carlo simulations and stochastic modeling in an effort to increase the chances of identifying robust trading strategies but these methods also have severe limitations due to changing market conditions, selection bias and data snooping. In this paper we present two examples that demonstrate the limitation of quantitative evaluation of trading strategies and we claim that the most effective way of guarding against overfitting and selection bias is by limiting the applications of backtesting to a class of strategies that employ similar but simple predictors of price. We claim that determining when market conditions change is in many cases fundamentally more important than any quantitative claims about trading strategy evaluation.

1. Introduction

Traders do not always have the luxury of testing trading strategies on a forward sample with real money because this takes time and it is costly in case of performance failure. Therefore, traders look for ex-ante measures of robustness of trading strategies and to achieve that resort to the use of quantitative analysis. Usually, a trading strategy is developed on in-sample data and validated on out-of-sample data. Although the academic community has been aware of the limited effectiveness of out-of-sample validation when multiple trials are involved, the practitioner community has been slow in recognizing these problems.

Three papers from the academic community, among several others, have recently increased the awareness of the practitioner community about backtest overfitting and multiple trials when developing and evaluating trading strategies. However, the results in these papers deal only with the part of the problem related to overfitting and selection bias but not with the more important problem of the effect of changing market conditions on genuine strategies that overfit on persisting market conditions.

Methods for adjusting the Sharpe ratio, called the haircut Sharpe ratio, to account for multiple trials, are discussed in Harvey and Liu (2015). As we shall demonstrate with an example, these adjustments cannot guard against Type I error (false discoveries) when a genuine strategy is used but there is a change in market conditions. In Bailey et al. (2014), a different method is presented for determining the minimum backtest length required to assess the risk of overfitting, as a function of the number of trials involved in the development of a strategy. This method also fails to address the important problem of changing market conditions, as further acknowledged in Bailey et al. (2015). Both methods do not deal with the major cause of failures of genuine strategies that are overfitted on favorable market conditions but then fail due to changing market conditions although they are of value in the case of strategies developed via machine learning.

In Novy-Marx (2016), a differentiation is made between pure selection and overfitting bias, and their combination thereof, in the case of multiple signals. Critical values of the T-statistic are offered in the cases of pure selection and combinations of signals, called the best *k*-of-*n* strategy, as a function of the number of signals considered. The paper offers critical T-statistic values for correcting for data-mining bias due to overfitting and selection bias. While the results in the paper are interesting, one drawback is that they are based on generating random signals with real stock data from January 1995 to December 2014. Combining the random signals yields significant strategies with high values of the T-statistic. However, traders are actually interested in how popular trading rules perform in forecasting out-of-sample returns. It is also not entirely clear from this paper how Type II error is affected by correcting for data-mining bias using the critical values obtained from combining random signals. Discarding strategies that have high probability to perform well is an opportunity cost. After all, the job of a trader is to trade, not to perpetually analyze and evaluate strategies.

An important contribution of the results in the aforementioned three papers is that they have raised awareness about the impact of overfitting, selection bias and data-snooping, especially during a period of time when there is renewed interest in machine learning applications to trading strategy discovery. However, practitioners have known for long that genuine trading strategies fail primarily when market conditions change because they cannot maintain positive expectation. One reason for the slow adoption of quantitative methods in strategy evaluation by practitioners is due to their limited effectiveness, especially when these methods become another metric to guide the strategy development

process. In such case, instead of minimizing data-mining bias, these quantitative methods actually become another factor that contributes to it.

In addition to the efforts by the academic community, practitioners of trading strategy evaluation have also attempted to deal with the problem of overfitting, selection bias and data-snooping with various ad-hoc quantitative methods.

In his popular book, *Evidence-based Technical Analysis*, David Aronson (2007) introduces the bootstrap and Monte Carlo permutation methods for generating sampling distributions used for statistical inference. In the case of the bootstrap, the null hypothesis is that the trading strategy mean return is 0 and, in the case of the Monte Carlo permutation, the null hypothesis is that the strategy possesses no intelligence in forecasting market returns. Aronson acknowledges that this approach is valid for independent trials and provides a set of heuristics for minimizing overfitting and selection bias, which are two components of data-mining bias (Harris, 2015). These heuristics involve limiting the number of tested rules, increasing sample size, considering correlated backtest results and limiting outliers and the variation of backtest results. However, these heuristics cannot limit the adverse impact from changing market conditions on genuine trading strategies, which is also the primary reason of failure.

Another method, called System Parameter Permutation (SPP), and its recent variant, System Parameter Randomization (SPR), was suggested by Walton (2014). This method involves applying a stochastic modeling approach for evaluating short-run and long-run performance estimates. The main advantage of SPP is that it does not rely on out-of-sample validation and this decreases data-snooping bias while it increases the power of the tests due to the larger sample. However, the first problem with SPP is that it requires selecting ex-ante a range of system parameters to subsequently vary and generate sampling distributions. This is problematic because data-mining bias does not only arise due to overfitting but also due to selection bias. In many cases selection bias is the main contributor to data-mining bias, for example when strategies have no parameters to vary. The second problem with SPP is that if it is repeatedly used under multiple trials, then it loses its effectiveness due to data snooping. The third and more serious problem is that all tests are conditioned on historical data and the probability of a Type I error is high under a change in market conditions. Therefore, SPP does not answer the following crucial question: How will strategy performance be affected if market conditions occur that are fundamentally different than those that were encountered during the analysis? As we shall see in Section 3.1 via an example, there is nothing SPP can do to determine a failure due to a massive change in market conditions.

2. A fundamental problem of trading strategy evaluation

There is a fundamental problem with quantitative trading strategy evaluation that renders it ineffective in most cases. Consider the following two sequences of returns: {0.1, 0.1, -0.1, 0.1, 0.1, -0.1, 0.1, 0.1, -0.1, 0.1}, {-0.1, -0.1, -0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1}. In both these cases, the win fraction is 0.7, the average return is 0.04 and the standard deviation is 0.0967. However, the maximum drop in equity for the first sequence is 10% but for the second sequence it is about 33%. A drop in excess of 20% to 25% could be a cause of abandoning a strategy or for the closure of a fund. Therefore, when analyzing trading strategies we are not only interested if a positive average trade, which is a simple average, will be achieved at some point in the future but whether time-averages will not result in excessive losses. The latter is a much more difficult problem, since trading strategy equity is path dependent, while realizing the true expectation of the population does not depend on the path but only on a sufficiently large sample of trades.

The fundamental problem of strategy evaluation is that the average of a sample converges to the expected value, or mean of the distribution, only at the limit of sufficient samples (Papoulis, 1965). This may be expressed as follows in the case of trading strategies

$$\lim_{N \to N_s} \frac{\sum_i^N R_i}{N} \to E[R] \quad (1)$$

where R_i are the values the random variable R takes, i.e., the strategy returns, N is the sample size and N_s is a sufficient sample size. The right side of equation 1 is the expectation, or mean of the return distribution, and the left side is the limit of the average return as the sample size becomes sufficient.

Due to non-stationary price series and fat-tailed distributions, a sufficiently large trade sample is required for determining the true expectation of the returns of a trading strategy and there is always high risk that this sample size will not be realized before uncle point and effective ruin. Statistical analysis allows using samples to estimate population properties. The problem with this approach in trading strategy evaluation is that due to changing market conditions and associated changes in distribution parameters, the significance of a strategy is a random variable while short-term performance depends on time-averages rather than on ensemble averages. These realities of trading system evaluation are often obscured in academic papers when in fact the idea is simple: trading strategies can get effectively ruined before their average return converges to the true mean of the distribution of returns. As a result, determining the statistical significance of a strategy based on limited samples far from suffices in assuring that it will continue generating profits (Harris, 2015). In fact, as we will show with two examples below that reflect two fundamental modes of behavior of market price series, momentum and mean-

reversion, Type I and Type II errors are typically quite large when market conditions change. Thus, the main message of this paper is that any claims from strategy evaluation that do not account for changing market conditions are severely limited because they ignore this most important cause of strategy performance deterioration.

3. Quantitative analysis limitations

In this section, we demonstrate the limitations of quantitative analysis by considering two genuine trading strategies. The first strategy, in Section 3.1, performed well for 48 years in the S&P 500 index but afterwards failed due to a change in market conditions. The second strategy, in Section 3.2, did not perform well in the same 48 year period but afterwards has outperformed the index on a both absolute and risk-adjusted basis. In Section 3.3 we attribute the changes in the performance of the two strategies to a change in markets conditions and specifically to a structural shift from positive to negative autocorrelation in daily returns. We argue that quantitative claims based on either statistical significance or stochastic modeling cannot account for the change in performance of these two strategies and are severely limited in that regard.

3.1. Time series smoothing example

Simple moving average smoothing is used in popular strategies that attempt to capture price trends. In academic papers these strategies are classified under price series momentum but practitioners use the more descriptive term *trend-following*. The general idea is that a strategy based on moving average smoothing can generate alpha by tracking the difference between price and a moving average. In the case of a long-only strategy, if the difference between price and the moving average is positive, then the strategy stays invested and goes flat if the difference turns negative.

Let P_t, t = 1, …, T, be time series observations. An m-month (simple) moving average (MA) at time t is computed as follows (Chen, 2013)

$$MA_t(m) = \frac{P_t + P_{t-1} + \cdots + P_{t-m+1}}{m}, \qquad t = m, \ldots, T \qquad (2)$$

Note that $MA_t(m)$ is the 1-step ahead forecast of P_{t+1} at time t.

The trading strategy is as follows

$$\text{Buy if } P_t - MA_t(m) > 0$$

$$\text{Sell if } P_t - MA_t(m) \leq 0$$

Trading strategies based on simple moving averages have been used for several decades by Commodity Trading Advisors (CTAs) and other market practitioners. These strategies

gained popularity for equity market timing after the dot-com bubble and especially after the financial crisis crash because of their potential of avoiding large drawdown caused by market crashes. A recent empirical study by Glabadanidis (2015) reported exceptional performance of moving averages strategies in the monthly timeframe and caused another round of renewed interest in these strategies by fund managers and also by developers of alternative products such as Exchange Traded Funds. However, Zakamulin (2016) has shown that the empirical study by Glabadanidis (2015) involved look-ahead bias and in reality the performance of market timing based on moving averages is indistinguishable from that of a buy-and-hold strategy.

Besides the look-ahead bias issue that was identified by Zakamulin (2016), there is another issue with the moving average market timing strategies in the empirical study by Glabadanidis (2015). Specifically, the timeframe compression from daily to monthly excludes two months of daily price changes from moving average calculations because the first difference available for the strategy is that between the monthly closing price P_t and $MA_t(2)$. Two months worth of daily data is about 64 trading days. This introduces selection bias because it filters out values of the moving average that have ceased to perform well due to changes in market conditions, as it will be shown next with an example, which is a long-only moving average strategy as defined above in the daily timeframe applied to S&P 500 index data downloaded from Yahoo! Finance from 01/03/1950 to 12/31/1997.

Period m	Annualized Return	Annualized Std Dev	Worst Drawdown	Annualized Sharpe ($R_f = 0$)	Sharpe Rank	T-stat
2	0.169	0.089	0.114	1.90	1	13.1635
3	0.148	0.088	0.142	1.68	2	11.6393
4	0.129	0.088	0.257	1.46	4	10.1151
5	0.125	0.089	0.165	1.41	6	9.76876
6	0.133	0.088	0.160	1.52	3	10.5308
7	0.126	0.087	0.145	1.45	5	10.0458
8	0.125	0.089	9.182	1.38	7	9.5609
9	0.108	0.087	0.204	1.25	9	8.6602
10	0.110	0.086	0.169	1.28	8	8.8681
11	**0.106**	**0.087**	**0.185**	**1.23**	**10**	**8.5216**
12	0.094	0.086	0.230	1.09	15	7.5517
13	0.099	0.086	0.244	1.14	11	7.8981
14	0.099	0.087	0.175	1.14	12	7.8981
15	0.095	0.087	0.192	1.10	13	7.6210
16	0.092	0.087	0.193	1.06	16	7.3438
17	0.096	0.087	0.210	1.10	14	7.6210
18	0.092	0.087	0.186	1.05	17	7.2746
19	0.091	0.087	0.202	1.05	18	7.2746
20	0.089	0.087	0.185	1.02	19	7.0667
Buy & hold	0.083	0.132	0.482	0.67	-	-

Table 1: Backtest results in S&P 500 daily data from 01/03/1950 to 12/31/1997 of a moving average strategy with period m values from 2 to 20, in increments of 1. Median Sharpe performance is shown in bold. The T-Stat is given as follows: $T\text{-}stat = Sharpe \times \sqrt{number\ of\ years}$.

The period m of the moving average is varied from 2 to 20 days. The available equity is fully invested when a buy signal is generated. The trade price is the closing price.

Commission is set at $0.01 per share[1]. The strategy goes flat when a sell signal is generated. The results of the backtest are shown in Table 1.

It is interesting to notice from Table 1 that the best performance (Sharpe = 1.90) is achieved for m = 2 and as m increases, performance decreases. The strategy outperforms buy and hold for all values of m between 2 and 20. T-statistic values indicate a highly significant strategy. It appears that the only question a strategy developer would face concerns which moving average period to choose in real application. Although the choice m=2 shows the best performance, this could be a random selection and a different value of m could perform better in the future. The Sharpe values are ranked and the period with median performance is chosen, which is m = 11. The corresponding t-statistic is 8.5216. This value of the t-statistic is high enough to accommodate most corrections for multiple trials and back-test overfitting, discussed in Section 1, and still indicate a significant result. However, also note that the 19 trials in Table 1 are not independent since the strategies are highly correlated and corrections to the T-statistic may not apply in the sense of Harvey and Liu (2014). In addition, a large trade sample limits backtest overfitting in the sense of Bailey at al. (2014) and significance is high.

Table 2 shows the forward performance of the strategy for m = 11. The annualized return is negative, worst drawdown is in excess of 55% and Sharpe is -0.13, i.e., the strategy fails in the forward sample.

Period m	Annualized Return	Annualized Std Dev	Worst Drawdown	Annualized Sharpe ($R_f = 0$)
11	-0.016	0.128	0.552	-0.13
Buy & hold	0.042	0.199	0.568	0.21

Table 2. Moving average strategy performance for m = 11 with S&P 500 daily data from 01/02/1998 to 5/31/2016.

In fact, the strategy generated negative performance in the forward sample for all values of m between 2 and 20. It is interesting to note that the value of m = 2 that generated the best performance in the in-sample, generated the worst performance in the forward sample. It is also interesting to note that the value of m = 11 that generated median performance in the in-sample, also generated median performance in the forward sample.

What could a practitioner have done near the end of 1997 to prevent using a strategy such as the one presented above that experienced a massive failure afterwards? Actually, there was little that could be done. This strategy had no chance to perform well due to a major shift in market conditions, as will be discussed in Section 4. No correction for significance and backtest overfitting would have provided an indication of the coming failure of this strategy. In addition, no analysis based on stochastic modeling could anticipate the major deterioration of performance. These methods for assessing the

[1] In the case of S&P 500 we consider hypothetical shares since indexes are not tradable directly.

significance and robustness of trading strategies have severe limitations in the sense that they do not deal with a major cause of failure, which amounts to changes in market conditions. The fact is that the size of a sufficient sample in Equation 1 is not known except if trivial assumptions are made about the distribution of returns. If market condition change in the future, then the results of statistical analysis are no longer valid.

A bootstrap of the in-sample trade returns of the moving average strategy with m = 11 generated the sampling distribution shown in Figure 1. The distribution was centered at zero mean. The standard deviation is 0.00895, kurtosis is 0.00124 and the skew is 0.032. As expected, the sampling distribution is close to normal by virtue of Central Limit Theorem[2]. Note that the mean return of the strategy for m = 11 is 0.00468 and it is more than 7 standard deviations away from zero mean. Therefore, we can reject the null hypothesis that the mean return of the system came from a distribution with zero mean.

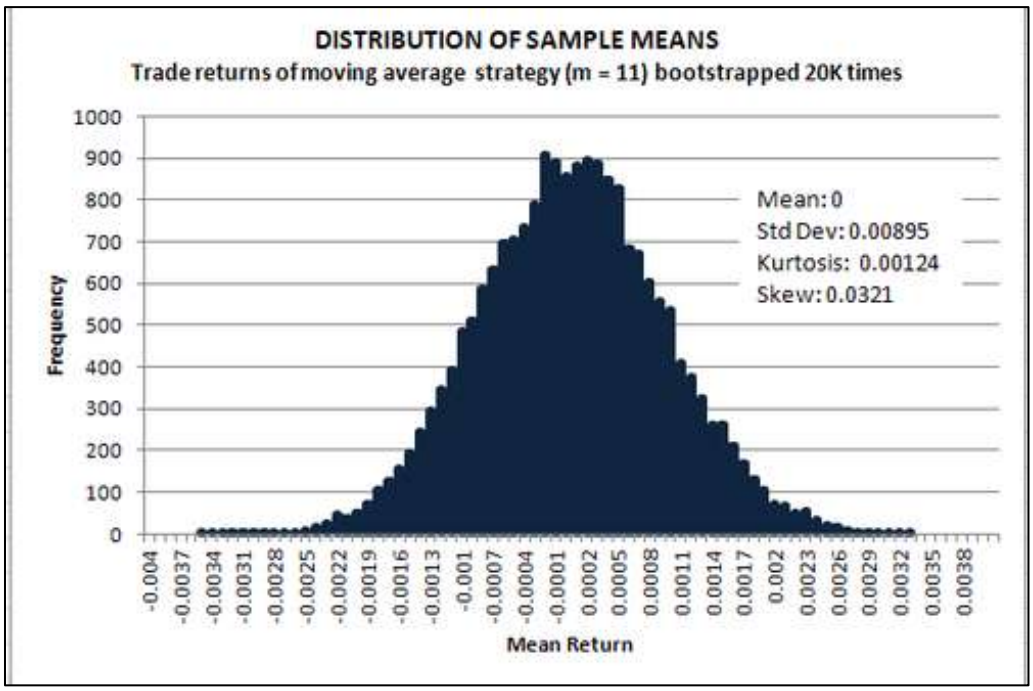

Figure 1: Sampling distribution of bootstrapped trade returns (with replacement) of the moving average system in Section 3.1with period m = 11. The distribution is centered at 0 mean.

The hypothesis test using in-sample results fails to anticipate the performance deterioration of the strategy in the forward sample. In fact, almost any conceivable statistical test, or analysis based on stochastic modeling, fails because the failure in the forward sample is not related to low statistical significance but to a massive change in market conditions, as it will be shown in Section 3.3.

[2] Note that the Central Limit Theorem does not apply in the case of all distributions. An example is the Cauchy distribution.

Figure 2 shows the minimum backtest length required to prevent skill-less strategies to be generated with a Sharpe ratio of 1, as a function of the number of trials, according to Bailey et al. (2014). The backtest length in the case of the moving average strategy considered in this section is about 48 years and only 19 trails were considered to choose the period m. Furthermore, the strategy configurations are not independent and the application of the results in Figure 2 is conservative. Therefore, Figure 2 may not even apply in this case. Furthermore, the claim that relates backtest overfitting to backtest length and number of trials is limited in the sense that for a large class of strategies, such as those that employ moving average smoothing, if market conditions do not change then a best fit to historical data may actually be desirable. It is only when market conditions change that a best fit has higher probability of failure in a forward sample.

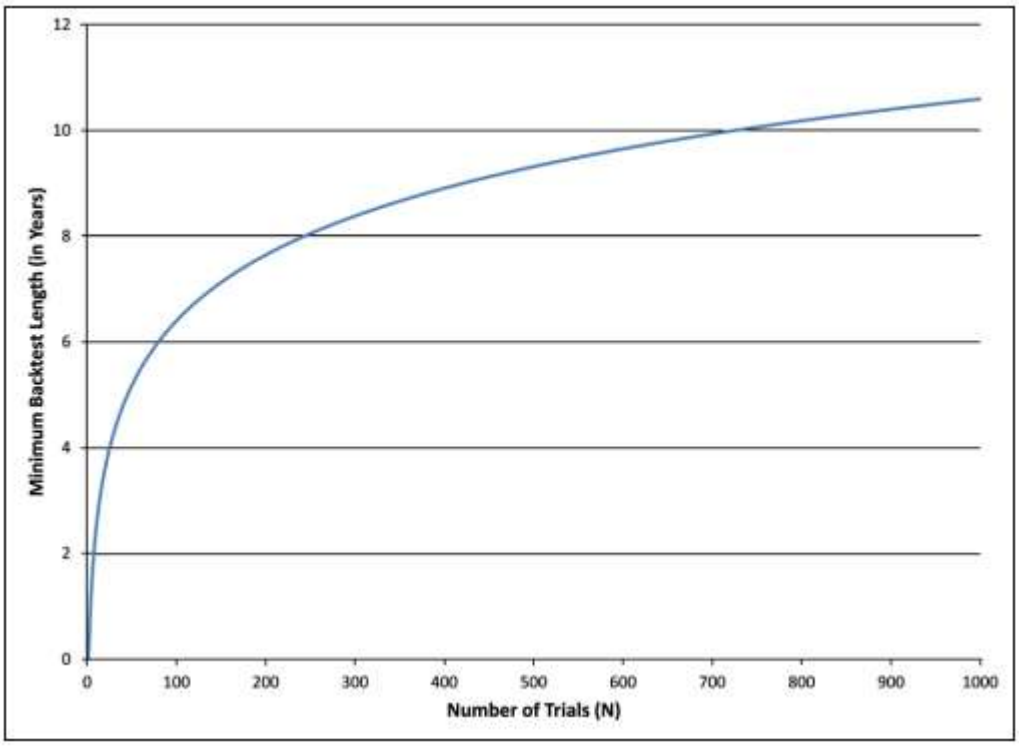

Figure 2: Required minimum Backtest Length in years for avoiding overfitting, as a function of the number of trials. Source: Bailey et al. (2014)

Figure 3 shows the empirical T-statistic values for correcting for data-mining bias in the case of strategies with multiple signals Novy-Marx (2016). As related to the moving average strategy considered in this section, it does not matter whether up to k=10 different strategy variations are combined based on the n = 19 different values of the moving average period m: all calculated T-statistics in Table 1 have values much larger than the critical values (yellow dotted and solid lines.) The critical value of the T-statistic starts a little above 3 and peaks at less than 5.5. The minimum value of T-statistic in Table 1 is 7.0667 for m = 20 and the maximum is 13.1635 for m = 2.

Regarding SPP tests based on stochastic modeling (Walton, 2014), again these have little or no application to the moving average example of this section. All 19 variations of the strategy are significant in the in-sample period of 48 years with high T-statistic values. In fact, for SPP to make sense there must be a sufficiently large number of parameters to vary. Most strategy developers avoid multiple parameters since the odds of overfitting are high. In essence, SPP attempts to solve a problem is requires to be present. However, most practitioners would like to attack the problem at its source and limit the number of free parameters in a strategy. This appears to be a much more sensible approach when it is possible to implement.

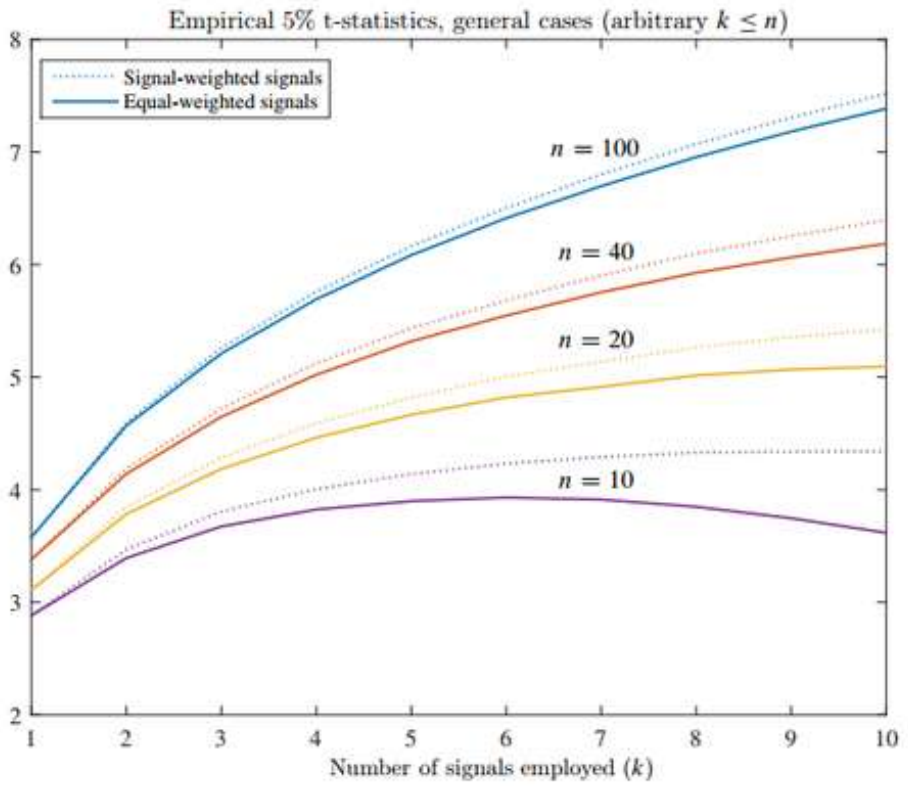

Figure 3: Empirical T-statistic values for correcting for data-mining bias in the case of strategies with multiple signals. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best k performing signals after considering n candidate signals. Solid lines show the cases when the composite signal is constructed by equal-weighting the k best performing candidate signals, and dotted lines the cases when the composite signal is constructed by signal-weighting the signals. Source: Novy-Marx (2016).

3.2 Mean-reversion strategy example

The term mean-reversion is a general term and can be described in several ways: For example, it could imply prices reverting to longer-term mean, or when a period of above

average returns is followed by a period of below average returns[3]. For some practitioners, mean-reversion is manifested in terms of overbought/oversold conditions that can be measured with popular indicators such as Relative Strength Index (RSI). Fama and Roll first discovered significant evidence of mean-reversion in individual stock returns over long time periods (Smith and Pantilei, 2013). In the example of this section, we consider a long-only strategy where two consecutive negative simple returns trigger a buy signal and two consecutive positive simple returns trigger a sell signal. Thus, after two consecutive negative simple returns the strategy expects a reversion to higher prices.

The simple return at period t is denoted by $R_{t,1}$ and defined as follows (Chen, 2013):

$$R_{t,1} = \frac{P_t}{P_{t-1}} - 1 \qquad (3)$$

where P_t is the price at period *t*. The trading strategy is as follows

Buy if $R_{t,1} \leq 0$ AND $R_{t-1,1} \leq 0$

Sell if $R_{t,1} > 0$ AND $R_{t-1,1} > 0$

The performance of the above trading strategy is next backtested in the daily timeframe with S&P 500 index data downloaded from Yahoo! Finance from 01/03/1950 to 12/31/1997. The available equity is fully invested when a new long position is signaled. The trade price is the closing price. Commission is set at $0.01 per share.[4] The strategy goes flat when a sell signal is generated. The results of the backtest are shown in Table 3.

	Annualized Return	Annualized Std Dev	Worst Drawdown	Annualized Sharpe ($R_f = 0$)
Strategy	**-0.038**	**0.098**	**0.742**	**-0.04**
Buy & hold	**0.083**	**0.132**	**0.482**	**0.67**

Table 3: Backtest results in S&P 500 daily data from 01/03/1950 to 12/31/1997 of the mean-reversion strategy where two consecutive negative simple returns trigger a buy signal and two consecutive positive returns trigger a sell signal. The strategy is unprofitable in the time period considered.

From the results in Table 3 it may be seen that the mean-reversion strategy was unprofitable in the tested period. The annualized return is negative and the maximum drawdown is in excess of 74%. Sharpe ratio is negative since this is a losing strategy.

Table 4 shows the forward performance in S&P 500 daily data from 01/02/1998 to 5/31/2016.

[3] Thanks to Valeriy Zakamulin for a discussion on the subject of mean-reversion

[4] In the case of S&P 500 we consider hypothetical shares since indexes are not tradable directly.

	Annualized Return	Annualized Std Dev	Worst Drawdown	Annualized Sharpe ($R_f = 0$)
Strategy	**0.070**	**0.151**	**0.323**	**0.47**
Buy & hold	**0.042**	**0.199**	**0.568**	**0.21**

Table 4: Backtest results in S&P 500 daily data from 01/02/1998 to 5/31/2016 of the mean-reversion strategy where two consecutive negative simple returns trigger a buy signal and two consecutive positive simple returns trigger a sell signal. The strategy is highly profitable in the time period considered.

The strategy annualized return outperforms that of buy and hold by nearly 300 basis point and worst drawdown is about 40% lower. Volatility is also lower and Sharpe is 0.47 for the strategy as compared to 0.21 for buy and hold. The T-statistic is 2.02. This means that the forward profitability is about two standard deviations away from null profitability and the significance level is roughly 5%.

Therefore, a strategy with negative performance is a period of 48 years outperformed buy and hold in the following 18.5 years. Application of quantitative analysis in the in-sample would have lead to a Type II error (missed discovery) and to opportunity loss.

There are more examples of simple return-reversion strategies that were unprofitable in the past in S&P 500 but have delivered high profitability in recent years due to a change in market conditions (Harris, 2016).

3.3 The impact of changing market conditions

In Section 3.1, we considered a strategy based on simple average smoothing that worked exceptionally well for 48 years only to fail in the following 18.5 years. Specifically, the strategy worked well in S&P 500 index from 1950 to 1997 but has failed thereafter. Quantitative analysis would have failed to predict the failure and would have resulted in a Type I error (false discovery.)

In Section 3.2, we considered a strategy with negative profitability in S&P 500 from 1950 to 1997 that reversed to outperformance thereafter. In this case quantitative analysis would have resulted in a Type II error (false rejection.)

In fact, the main problem faced by practitioners of trading strategy development is not with the correct application of quantitative analysis but with determining when market conditions will render a previously robust strategy unprofitable. In the case of the particular strategies considered in Sections 3.1 and 3.2, a better understanding of the cause of the change in profitability may be obtained by looking at the rolling 252-day autocorrelation of S&P 500 daily simple returns from 1950 to 2016, shown in Figure 4.

Specifically, autocorrelation was overall positive from 1950 to 1997 and also significant from about 1966 to 1988. However, the autocorrelation became non-significant afterwards and mostly negative. That caused a major change in market dynamics that

impacted the profitability of certain trading strategies. These strategies cannot be called overfitted; they were strategies that exploited particular dynamics of the market.

Autocorrelation is only one particular property of price series. There are many other properties that can affect strategy performance. Some of those properties, and features in general, are not easily modeled.

Figure 4: Rolling autocorrelation of simple returns for a 252-day period. The horizontal bold lines are the 95% confidence bands and the vertical bold line marks a structural change that occurred near the end of 1997. Specifically, after about 1997, autocorrelation falls below the upper significance band and it is mostly negative.

Figure 5 shows the performance of the moving average strategy discussed in Section 3.1 but on monthly data of the SPDR S&P 500 ETF (SPY) from inception to 07/13/2016 with the period m set to 10 months. Table 5 lists key performance parameters. The strategy has outperformed buy and hold by avoiding the large drawdown levels induced by the dot com bubble crash and financial crisis. Specifically, annualized Sharpe for the strategy is 0.939 as compared to 0.607 for buy and hold. This type of "timing model" has had outstanding performance but Zakamulin (2016) has shown that its performance is (statistically) indistinguishable from that of buy and hold. The fact is that this and related strategies are (over)fitted on specific market conditions and features of price series. Specifically, they rely on "macro V-bottoms", such as those formed after the dot com and financial crisis bear markets and on extended trends thereafter with relatively low volatility. If these conditions are not met, these strategies may accumulate significant losses. An example is the performance of the same strategy in iShares MSCI Emerging Markets ETF (EEM) in the period 01/2009 to 06/2016, shown in Figure 6. An extended

sideways market since 2012 has caused the strategy to generate negative returns and a higher drawdown as compared to buy and hold. When using this and related strategies, practitioners are more interested in whether market conditions will remain favorable and not so much in their quantitative evaluation. Should practitioner abandon these strategies simply because quantitative analysis indicates so? The answer to this question is not easy and this is only one reason that profiting from market moves is not an easy endeavor. Deciding on the right strategy based solely on the results of quantitative analysis may be introducing a high opportunity cost due to Type II error. Practitioners believe that if a specific market dynamic has persisted for a long time it will continue doing so in the future. Specifically here we are referring to "price series momentum premium" that can be harvested by moving averages of longer periods. Although as mentioned in Section 3.1, this particular choice of moving average periods entails selection bias, it is not entirely clear that quantitative analysis can provide answers to the above question.

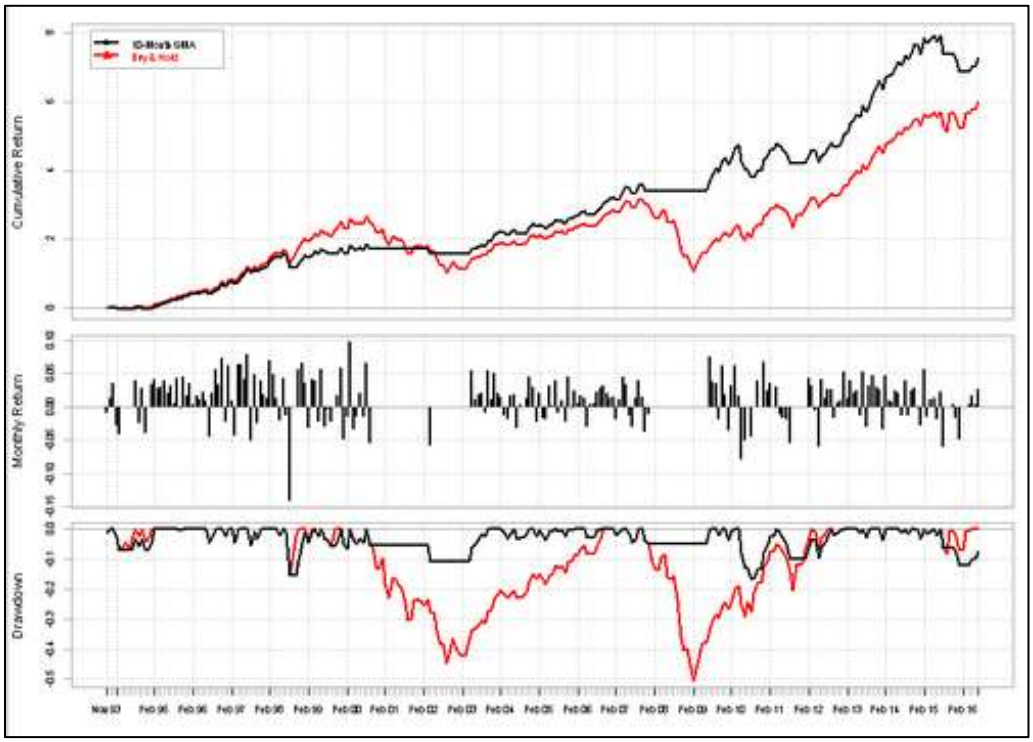

Figure 5: Performance of the moving average strategy in Section 3.1 on monthly SPY data with moving average period m set to 10 months. The strategy timing allows avoiding the large drawdown levels of the dot com and financial crisis bear markets, resulting in high risk-adjusted performance as compared to buy and hold. R code courtesy of Ilya Kipnis, quantstrattrader.wordpress.com

	10-Month SMA	Buy & Hold
Annualized Return	0.0974000	0.0893000
Annualized Std Dev	0.1037000	0.1470000
Annualized Sharpe (Rf=0%)	0.9388000	0.6074000
Worst Drawdown	0.1663486	0.5078482
Calmar Ratio	0.5852313	0.1757602

Table 5: Performance metrics of the strategy in Figure 5. Sharpe for the strategy is 0.938 versus 0.607 for buy and hold, indicating significant outperformance.

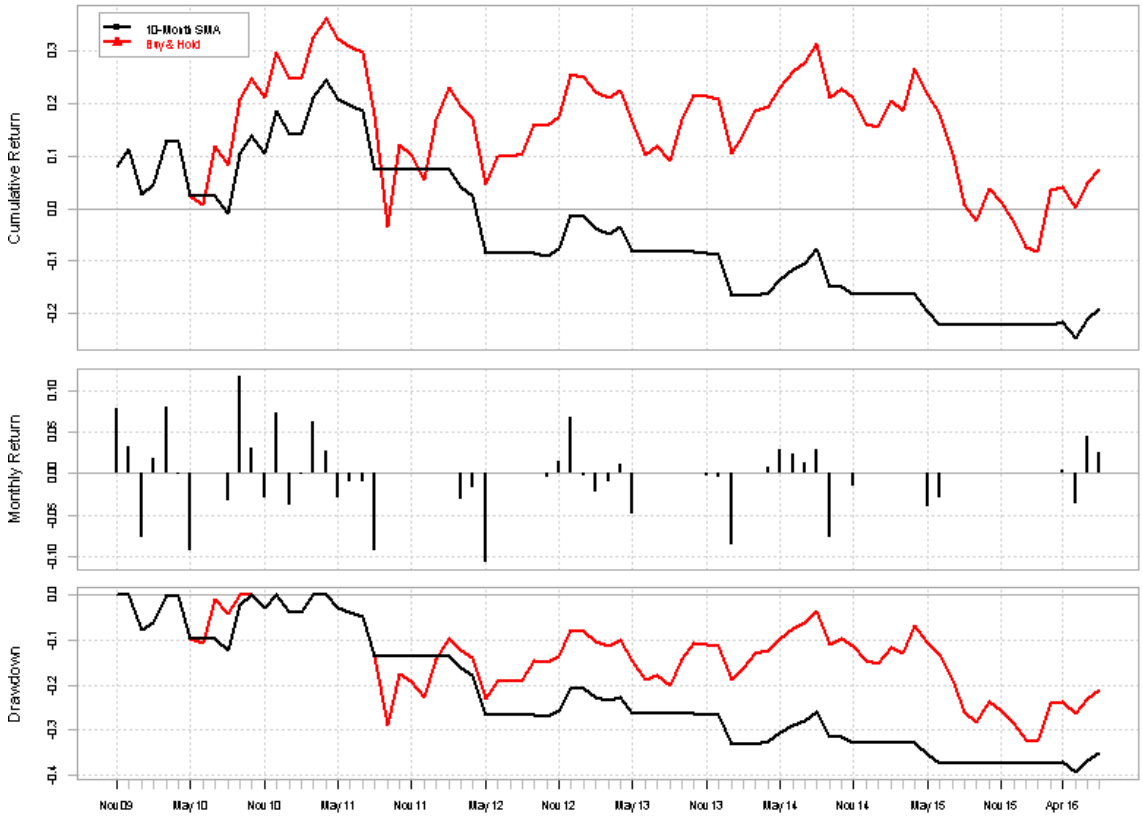

Figure 6: Performance of the moving average strategy in Section 3.1 but on monthly EEM data with moving average period m set to 10 months. The strategy fails due to prolonged sideways market activity. R code courtesy of Ilya Kipnis, quantstrattrader.wordpress.com

While the academic community is infatuated with statistical significance, possibly because it offers a foundation for producing papers that provide credit towards tenure, practitioners are more interested in whether specific strategies will continue working in the future, i.e., whether market conditions will remain favorable for these strategies. It is a secondary issue for practitioners whether these strategies are the outcome of selection bias and data-snooping, as long as they have worked for extended periods of time. For example, the moving average smoothing strategy in Figure 5 has worked well for 23 years and the question practitioners would like to answer is whether the same conditions that allowed the strategy to generate high risk-adjusted returns will be present in the future. If the strategy fails after one year then statistical significance analysis has merit. But if it will continue working well for the next 23 years, then a practitioner can generate

enough wealth for his account and client accounts. In fact, some practitioners may even perceive statistical analysis as a strawman, i.e., an artificial problem for the academic community to attack. The truth is somewhere in between: statistical analysis is useful, especially in data-mining, but understanding the nature of market conditions that contribute to profitability and when they are about to change is also as important, and in some cases, even more important.

5. Conclusion

Quantitative analysis of trading strategies is useful, especially in raising awareness among market practitioners about the perils of multiple trails and backtest overfitting. However, this applies more to a specific class of practitioners that use backtesting in naïve ways. One such naïve way is the application of machine learning and neural networks to trading strategy discovery with a large variety of technical indicators, such as for example, various types of moving averages, price oscillators, etc. In this case quantitative analysis provides a framework for evaluating the significance of the results in the context of multiple trials and backtest overfitting. However, most professional practitioners already know that multiple trials can be an exercise in futility unless certain conditions are met. For example, Leda Braga has stated the following during an interview[5]:

"There's a creative moment when you think of a hypothesis, maybe it's that interest rate data drives currency rates. So we think about that first before mining the data. We don't mine the data to come up with ideas."

Most professional practitioners either follow Braga's approach or when they use machine learning and data-mining they make sure to abide by Aronson's heuristics (Aronson, 2007). Only naive practitioners feed data to a machine learning model in hope that it will generate a significant result. Quantitative analysis shows that results from multiple trials can be misleading. However, quantitative analysis of trading strategies that are based on sound models of price behavior and are even (over)fitted cannot answer when these models will stop working, as that can occur after one week or even after several decades. Therefore, quantitative analysis has certain limitations and cannot effectively deal with the issues faced by all trading strategy developers. Trading strategy development will always remain an empirical field that is partly science but also partly art. Although the academic community has contributed significantly in raising awareness about certain issues, it cannot provide a framework for generating those "creative moments" Leda Braga referred to above but only investigate whether a moment was not as creative as was expected. Although this is partly progress, it is far from a solution to the problem, if such a solution exists at all.

Finally, the only effective way of minimizing bias when developing strategies is by limiting the use of backtesting to a small class of models that have similar characteristics so that multiple comparisons are minimized and correlation between strategies is high.

[5] http://www.bloomberg.com/news/articles/2015-02-26/leda-braga-s-bluetrend-delivers-a-top-hedge-fund-performance

Essentially, bias arises from three processes: optimization of parameters, selection/rejections of models and reuse of data. Optimization can result in overfitting to noise, selection/rejection introduces survivorship bias and data reuse introduces data-snooping bias. The combination of these biases is what is usually referred to as data-mining bias. Trading strategy development is a complex process that is path dependent and there are no general quantitative solutions. One sound approach in dealing with the complexity of this process and with the potential of being fooled by random results is by limiting its application.

References

Aronson, D. R. (2007), Evidence-based Technical Analysis: Applying the Scientific Method and Statistical Inference to Trading Signals, John Wiley & Sons, Inc., Hoboken, NJ

Bailey, D., J. Borwein, M. López de Prado and J. Zhu (2014), "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

Bailey, D., J. Borwein, M. López de Prado and J. Zhu (2015), "The Probability OF Backtest Overfitting", Journal of Computational Finance. Available at: http://ssrn.com/abstract=2308659
Chen, Mey-Yan (2013), "Financial Time Series and Their Characteristics", National Chung Hsing University, Available at http://web.nchu.edu.tw/~finmyc/tim_serp.pdf

Glabadanidis, P. (2015). "Market Timing With Moving Averages", *International Review of Finance, 15* (3), 387–425.

Harris, M (2015), Fooled by Technical Analysis: The perils of charting, backtesting and data-mining, Price Action Lab. Available at http://www.priceactionlab.com/Blog/the-book/

Harris, M. (2016), "Performance Of Three Mean Reversion Strategies", Price Action Lab. Available at http://www.priceactionlab.com/Blog/2016/06/performance-mean-reversion-strategies/

Harvey, C. and Y. Liu (2015), "Backtesting" Working paper, Duke University. Available at http://ssrn.com/abstract=2345489

Novy-Marx, R. (2016). "Testing strategies based on multiple signals", Working paper. Available at http://rnm.simon.rochester.edu/research/MSES.pdf

Papoulis, A. (1965) Probability, Random Variables, and Stochastic Processes. McGraw-Hill, NY

Smith, D. and Pantilei S. (2013), "Do 'Dogs of the World' Bark or Bite? Evidence from Single-Country ETFs," Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2279246

Walton, D. (2014), "Know Your System! – Turning Data Mining from Bias to Benefit Through System Parameter Permutation", *2014 NAAIM Wagner Award Winner*, Available at http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2423187

Zakamulin, V. (2016). "Revisiting the Profitability of Market Timing with Moving Averages", Working paper, University of Agder. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2743119