Systematic Testing of Systematic Trading Strategies

Kovlin Perumal[†] and Emlyn Flint[††]

13th March 2018

Abstract

Systematic trading is a method that is currently extremely popular in the investment world. The testing of systematic trading rules is usually done through back-testing and is at high risk of spurious accuracy as a result of the data-mining bias (DMB) present from testing multiple rules concurrently over the same history. The eradication of this DMB with the use of statistical methodologies is currently a relevant topic in investment research, illustrated by papers written by Chordia *et al.* (2017), Harvey and Liu (2014), Novy-Marx (2016) and Peterson (2015). This study collectively reviews the various statistical methodologies in place to test multiple systematic trading strategies and implements these methodologies under simulation with known artificial trading rules in order to critically compare and evaluate them.

Keywords: data-mining bias, systematic trading, artificial trading rules, backtesting, White's Reality Check, Monte Carlo Permutation, familywise error rate, false discovery rate, Hansen's Test for Superior Predictive Accuracy, Corradi & Swanson's Extension, Step-M method

[†]University of Cape Town, kovlinpermal@gmail.com

[††]Peregrine Securities and Department of Actuarial Science, University of Pretoria, emlyn.flint@gmail.com

The views expressed are our own and any errors made are our responsibility.

1 Introduction

Systematic trading is a method that is currently extremely popular in the investment world. It is a process whereby a buy or sell signal is generated from a rules-based quantitative process with the aim of meeting an investor's objectives given a defined constraint set. These systematic trading rules are evaluated through the method of backtesting, which involves the historic simulation of an algorithmic investment strategy (Bailey *et al.*, 2014).

However, evaluation in this way is at high risk of spurious accuracy as a result of the data-mining bias (DMB) present from testing multiple rules concurrently over the same history. This makes it difficult to distinguish between genuinely superior rules and false discoveries that happen to get lucky over the particular backtest. This problem results in resources being put into trading rules which perform well in-sample (in the backtest) but are unprofitable out-of-sample (in the real world).

DMB is eradicated through the use of various statistical methodologies. Solving the problem of DMB through the implementation of these statistical methodologies is currently very relevant. This is illustrated by the large number of studies into the subject in recent years. Examples of such research include papers by Chordia *et al.* (2017), Harvey and Liu (2014), Novy-Marx (2016) and Peterson (2015). These authors advocate the use of these statistical methods to eradicate DMB and will serve as the inspiration for our research.

Chordia *et al.* (2017) highlights the problem associated with data mining (or 'p-hacking') by practically evaluating 2.1 million trading strategies based on empirical fundamental stock data, using a variety of multiple hypothesis tests. These tests included the Bonferroni correction, White's Reality Check, Controlling the False Discovery Rate and various iterations of the Step-M method. The authors found that, although these strategies performed well according to conventional tests, very few strategies survived the multiple hypothesis evaluation. The surviving strategies were also found to have no apparent theoretical underpinnings according to the various additional economic hurdles employed.

Novy-Marx (2016) concludes that systematic multi-signal trading strategies cannot be evaluated using conventional tests. Through the conventional evaluation of multi-signal strategies, selected using purely random signals they show that DMB makes it very easy to produce strategies with excellent backtested performance that, by construction, have no power. Novy-Marx derives different test statistics that he suggests be used in order to remove the effects of DMB.

Harvey and Liu (2014) practically displays the ease at which seemingly profitable strategies can be produced by exploiting the effects of DMB with the use of multiple illustrative examples. They conclude that researchers must be aware of the problems associated with DMB and make use of multiple hypothesis testing frameworks. In particular, Harvey and Lui suggest that controlling the Family-wise Error Rate and False Discovery Rate should be used.

Our research provides a unique perspective on the matter by moving away from highlighting the problem of DMB and using empirical examples of trading strategies. Instead, our study focuses on critically evaluating the statistical methods used to eradicate DMB, with the aim of discovering which method performs best in the context of systematic trading strategies. To achieve this goal, rather than using empirical example strategies, we make use of artificial trading rules, simulated with controlled variables. This approach makes it possible to analyse the effects of each simulated market and strategy variable on the magnitude of DMB observed, as well as test the statistical methods over a wide range of market and strategy scenarios.

A contradictory paper by Harris (2016) states that this type of quantitative evaluation is obsolete. Harris claims that studying drastic market condition changes and the time at which they occur is fundamentally more important. Harris rationalises this by stating that these market changes cause the largest gains, the largest losses and the most invalidations with regard to trading strategies. This may be true to a degree however, statistical methods can still be used to identify profitable trading strategies in current market conditions. These conditions can be assumed to stay constant over short terms as drastic market changes occur infrequently. The underlying market characteristics which allow for strategy profitability can then be identified through analysis and monitored for any change. This shows that quantitative claims made through the use of statistical methodologies still have a meaningful use.

The remainder of this study is set out as follows. Section 2 provides an overview of the market and strategy simulation framework, as well as the statistical methods to be tested. Section 3 briefly describes the simulation process. The results of the study are presented in Section 4 and can be split into two parts. The first part focuses on the effect of each simulation parameter on the total magnitude of DMB observed and the second presents the general performance of each statistical method across various market and strategy scenarios. Section 5 concludes the study.

2 Methodology and Data

The tests performed in this paper involve simulation of both artificial market data and artificial trading rules (ATRs) with known combinations of parameters. The general effects of changing these parameters on the magnitude of data-mining bias is measured. The specific effects of changing these parameters on relative statistical method performance is also measured. The performance of these statistical methods is then evaluated by comparing the number of false discoveries made by each method. The average p-values obtained from each method are observed against strategy profitability to give an indication of relative power. This p-value refers to the probability that the mean return produced by the simulated ATR for a simulated market path and a single combination of parameters, is equal to zero. The artificial market, artificial

trading rules and parameters associated with each will now be outlined.

2.1 Market Data Simulation

In order to realistically simulate market data, the Merton Jump Diffusion model is selected as the simulation model. This model captures two key features of real world markets deemed important, namely the variance of market returns and the presence of jumps. Table 1 provides an overview of the selected market parameter ranges.

Table 1: Market simulation parameters

Parameter	Symbol	Possible Values
Drift of Diffusion	u_d	0
Diffusion Standard Deviation	σ_d	0.1, 0.3, 0.5
Poisson Arrival Rate	λ	0, 0.1, 0.2
Drift of Jump	u_j	0
Jump Standard Deviation	σ_j	0.01
Time Step (months)	dt	1 (in months)
Number of Paths Simulated	N	100

2.2 Artificial Trading Rule Simulation

ATRs are simulated in order to generate mean returns for each simulated market path. These ATRs are defined according to a probability of success p, which is simply the probability of the ATR producing a signal that will match the sign (positive or negative) of the market return on the applicable time step. Furthermore, the number of ATRs simulated for testing and length of backtest (strategy length) are also controlled parameters. Table 2 provides an overview of the selected ATR parameter ranges.

2.3 Statistical Methods

After simulating the market data and ATRs, the mean return for these strategies are calculated. The underlying question posed is: "are the mean returns significantly greater than zero?".

Table 2: Artificial trading rule parameters

Parameter	Symbol	Possible Values
Probability of Success	p	0.25, 0.4, 0.5, 0.6, 0.75
Number of Rules Simulated	K	2, 10, 25, 50, 100
Strategy Length (months)	T	12, 60, 120, 180

The answer to this question is susceptible to the effects of data-mining bias. Since we know (through controlling the ATR success parameter p) whether these strategies will be profitable or not, we can show how data-mining bias presents itself in conventional t-tests and the extent to which it is eradicated using a variety statistical methods. These statistical methods will now be briefly described[2].

2.3.1 Controlling Family Wise Error Rate (FWER)

The key problem in multiple hypothesis testing is that the the probability of making a false discovery diverges from the initially defined α according to the number of hypotheses K tested[3]. This is known as data-mining bias.

The Family-Wise Error Rate (FWER) is defined as the probability of making at least one false discovery when testing multiple hypotheses (Harvey and Liu, 2013). Controlling the FWER is one of the fundamental tools needed to ensure that the number of falsely identified significant strategies be can be bounded. The simplest way of controlling the FWER is known as the Bonferroni correction. This states that in order to preserve α as the actual type 1 error rate (the error rate of making a false discovery) we should set $\alpha^* = \alpha/K$ and use α^* as the critical value (significance level) when performing each individual hypothesis test.

2.3.2 Controlling the False Discovery Rate (FDR)

The False Discovery Rate (FDR) is defined as the expected proportion of falsely rejected null hypotheses (Harvey and Liu, 2013). The principle way to control the FDR is by using the Benjamani & Hochberg Procedure (Benjamini and Hochberg, 1995). This procedure determines an upper bound for the number of false discoveries produced. The procedure will now be formally outlined[4].

Consider having already conducted the individual hypothesis tests $H_1, H_2, ..., H_k$ comparing our K strategies to the zero-mean benchmark and obtaining the p-values associated with each denoted by $p_1, p_2, ..., p_k$. Let $p_{(1)} \leq p_{(2)}, \leq ... \leq p_{(k)}$ be the ordered p-values for the corresponding hypothesis test $H_{(1)}, H_{(2)}, ..., H_{(k)}$. The Benjamani and Hochberg procedure states that in order to control for FDR at significance level α, reject all $H_{(i)}$ for which $p_{(i)} \leq \frac{(i)}{K}\alpha$. The largest rejected $p_{(i)}$ is then referred to as the Benjamini & Hochberg threshold and rejecting hypotheses according to this threshold ensures that $\mathbb{E}[\frac{\text{False Discoveries}}{\text{All Discoveries}}] \leq \alpha$, and thus $\mathbb{E}[\text{False Discoveries}] \leq \alpha\mathbb{E}[\text{All Discoveries}]$. In theory, this allows one to set the amount of false discoveries to a level which leaves sufficient power to find other significant results.

[2]See Appendix A1 for more mathematical detail.

[3]For an illustrative example on this divergence see Appendix A.1.1.

[4]For the motivation behind the proposal of this test by Benjamini and Hochberg (1995) see Appendix A.1.3.

2.3.3 White's Reality Check (WRC)

White's Reality Check (WRC) is the first comprehensive testing methodology implemented (White, 2000)[5]. The goal of WRC is to provide a multiple hypotheses testing method that moves away from a reliance on bounds, such as the Bonferroni correction, and that directly delivers appropriate p-values. White states that the null hypothesis tested should be,

$$H_0 : \max_{k=1,...,K} E(s(u_{k,t}) - s(u_{0,t})) \leq 0$$

$$H_A : \max_{k=1,...,K} E(s(u_{k,t}) - s(u_{0,t})) > 0,$$

where $s(.)$ is a defined 'satisfaction' function, $u_{k,t}$ is the return produced by strategy k at time t and $u_{0,t}$ is the benchmark return at time t. In the context of our practical implementation, we define $s(.)$ to simply be the mean strategy return and the benchmark is set to a mean return of zero. In alternate implementations of WRC, $s(.)$ can represent any performance metric (such as a Sharpe Ratio, Calmar Ratio, VaR etc). Intuitively this null hypothesis states that the best strategy encountered over a particular search has no superiority over the benchmark strategy with a mean return of zero. A further proposition in White (2000) states that for $t = 1, ..., T$,

$$\max_{k=1,...,K} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} ((s(u_{k,t+1}) - s(u_{0,t+1})) - E(s(u_{k,t}) - s(u_{0,t})) \rightarrow \max_{k=1,...,K} Z_k,$$

where Z_k is distributed $N(0, \mathbb{V}ar(s(u_k) - s(u_0))$. Building from this, the corresponding test statistic is given by,

$$TS_k^{WRC} = \max_{k=1,...,K} (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (s(u_{k,t+1}) - s(u_{0,t+1}))).$$

The distribution for the maximum of a Gaussian process is difficult to determine and is not a Gaussian process. As a result, the hypothesis for WRC is tested by generating a sampling distribution through the methods of bootstrapping. p-values for WRC are then attained from this sample distribution. The implementation of this method closely follows the algorithm outlined by Aronson (2011).

[5]White (2000) extended the test developed by Diebold and Mariano (1995) to allow for comparisons between multiple hypotheses. A description of this test is available in Appendix A.1.4.

2.3.4 Hansen's Test for Superior Predictive Ability (SPA)

Hansen (2005) proposes a new test for checking whether a model has "Superior Predictive Accuracy" (SPA) over a benchmark model. Hansen's test is implemented in a similar way as WRC but utilises a different test statistic and a sample-dependent distribution. Hansen defines the following studentised test statistic,

$$TS_k^{SPA} = \max[\max_{k=1,\ldots,K} \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(s(u_{k,t+1}) - s(u_{0,t+1}))}{\hat{w}_k}, 0],$$

where $\hat{w}_k^2$ is some consistent estimator of $w_k^2 = \mathbb{V}ar(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(s(u_{k,t+1}) - s(u_{0,t+1}))$. The estimator used in the practical implementation in this paper is $w_k^2 = B^{-1}\sum_{b=1}^{B}(T^{\frac{1}{2}}s(u_{k,t+1}) - T^{\frac{1}{2}}s(u_{0,t+1}))$ where B is the number of resamples used in the bootstrapping process. p-values for each TS_k^{SPA} are then obtained from a sample distribution generated through bootstrapping.

2.3.5 Monte Carlo Permutation (MCP)

The Monte Carlo Permutation (MCP) method, developed by Masters (2006), is based on the idea of testing whether an 'informed' model is significantly superior to a 'no-skill' model that is devoid of any predictive power. This method is implemented in our context according to the algorithm outlined by Aronson (2011). Specifically, by using bootstrapping to generate the sample distribution of a 'no-skill' rule's backtested performance. This 'no-skill' rule is created by randomly pairing the monthly returns of the simulated market with the ordered time series representing the sequence of ATR output values. This random pairing of rule values to market returns eliminates any predictive power the rule may have had and creates a 'no-skill' rule, which Aronson refers to as a 'noise rule'. The sample distribution of maximum mean returns produced by this noise rule is generated though bootstrapping and the mean return produced by each ATR tested against this to generate p-values.

2.3.6 Corradi & Swanson's Extension (CS)

Corradi and Swanson (2011), building on the work of White (2000), introduced a novel testing approach in which forecast combinations are evaluated through the examination of the quantiles of an expected loss distribution. In more detail, the models are compared by looking at cumulative distribution functions (CDFs) of prediction errors, for a given loss function, and the model whose CDF is stochastically dominated[6] is chosen to be the best performing.

The definitions originally used by Corradi and Swanson are converted for our application of testing trading strategies. A notable difference in our context is

[6]See Appendix A.1.2 for a full definition of first order stochastic dominance.

that we are looking for the rule that stochastically dominates the other rules in terms of our satisfaction function s, whereas Corradi and Swanson were looking for a rule that was stochastically dominated in terms of a loss function g. In order to use their methodology without major changes we define $g = -s$. Minimising the "loss" function g will thus be the same as maximising s. We note our usual definitions for $u_{k,t}$ and $s(u)$ and let $F_{g,k}(x)$ be the empirical distribution of $g(u_{k,t})$ evaluated at x and $\hat{F}_{g,k,T}(x)$ be its sample analog, i.e.,

$$\hat{F}_{g,k,T}(x) = \frac{1}{T}\sum_{t=1}^{T} 1_{\{g(u_{k,t})\leq x\}}.$$

The corresponding hypotheses are:

$$H_0 : \max_{k>0} \inf_{x\in X} (F_{g,0}(x) - F_{g,k}(x)) \geq 0$$

versus

$$H_A : \max_{k>0} \inf_{x\in X} (F_{g,0}(x) - F_{g,k}(x)) < 0.$$

Consider the case of a single strategy ($k = 1$) versus a benchmark strategy ($k = 0$). If $(F_{g,0}(x) - F_{g,1}(x)) \geq 0$ for all x, then the CDF associated with the benchmark strategy always lies above the CDF of the competing strategy. Then $g(u_{0,t})$ is (first order) stochastically dominated by $g(u_{1,t})$ and the benchmark is preferred. The dominated strategy is preferred because we are dealing with losses, which we would like to minimise. Alternatively, if we reject the null it implies that either rule 0 stochastically dominates rule 1 or the CDFs cross and further analysis is required to select the best performing rule.

Corradi & Swanson utilise the following test statistic when testing the null and alternative hypotheses outlined above,

$$L_{s,T} = -\max_{k>0} \inf_{x\in X} \sqrt{T}(\hat{F}_{g,0,T}(x) - \hat{F}_{g,k,T}(x))^{7}$$

Bootstrapping is then used to construct an empirical distribution and compare the calculated sample test statistic against the percentile of this empirical distribution.

2.3.7 Romano & Wolf's Test (Step-M)

Romano and Wolf (2005) propose a stepwise multiple testing procedure (also refered to as the 'Step-M' method) that asymptotically controls the family wise error rate, with the use of studentisation where applicable. Romano & Wolf's

[7] The negative sign in front of the statistic ensures that the statistic does not diverge under the null hypothesis.

method is thus a stepwise extension of WRC proposed in order to increase the power of the test, while still controlling the FWER at a given level α. The first step of the Step-M method is analogous to that of the WRC test. Let $w_{T,k}$ be a basic test statistic that will be used to test these hypotheses. Firstly, obtain K such test statistics, one for each strategy being tested against the benchmark. Then order these statistics and re-label them running from largest to smallest with the subscript (1) to (K) such that $w_{T,(1)} \geq w_{T,(2)} \geq ... \geq w_{T,(K)}$ and the corresponding hypotheses are, $H_{(1)}, ..., H_{(K)}$. Individual decisions are then made for each $w_{T,(k)}$ in a stepwise manner. For the first step, a rectangular joint confidence interval with nominal joint coverage probability $1 - \alpha$, for a chosen α is constructed. This confidence region is of the form:

$$[w_{T,(1)} - c_1, \inf) \times ... \times [w_{T,(K)} - c_1, \inf),$$

where c_1 is chosen to ensure the selected joint coverage probability, $1-\alpha$. This ensures that the FWER is maintained at α. If a particular individual confidence interval $[w_{T,(k)} - c_1, \inf)$ does not contain zero, then the corresponding null hypothesis $H_{(k)}$ is rejected. The stepwise component of this method now is realised. After m hypotheses are rejected in the first step, a second rectangular joint confidence interval is constructed for the remaining hypotheses, again with the nominal joint coverage probability of $1 - \alpha$. This confidence interval is of the form:

$$[w_{T,(m+1)} - c_2, \inf) \times ... \times [w_{T,(K)} - c_2, \inf),$$

where c_2 is selected to ensure the joint coverage probability of $1-\alpha$. This process is repeated until no further hypotheses can be rejected and one is left with a pool of significant strategies for which the FWER is controlled at α. Romano and Wolf (2005) detail the method of finding the c_i values exactly, the details of which are beyond the scope of this study. In our practical implementation the c_i values were obtained as the critical values of the sample distribution produced after bootstrapping at each step. This method of implementation is detailed in Chordia *et al.* (2017)

3 Testing Algorithm

The test carried out in this paper involves the simulation of a carefully controlled market environment with all parameters known at the outset. A sample path with a specific parameter set $\{u_d, \sigma_d, \text{ and } \lambda\}$ is simulated and K ATRs of length T with a specified probability of success p are generated. The 'satisfaction' for each strategy is then calculated as an indicator of strategy performance. This is simply the ATR's mean return over the simulated backtest.

The ATR with the best mean return is selected for testing and a new market path is simulated. This is repeated N times such that N maximum mean

returns and corresponding strategies are obtained for testing. This is done to simulate (N times) the real-world situation of investors choosing the "best-of-K" strategy without accounting for DMB. The testing of N maximums, as opposed to N observed means of varying profitability, also serves to stress test the statistical methods themselves in order to observe their performance in extreme situations.

The significance of these N "best-of-K" mean returns are then tested with M different statistical methods. The null hypothesis for every statistical method implemented states that the mean return produced by each strategy is equal to zero. The p-values obtained through the testing give the probability that this null hypothesis is true. A p-value > 0.05 indicates that a mean return is not significantly different to zero and a p-value < 0.05 indicates the opposite.

4 Results

In total, the results obtained after implementation were 720 000 p-values (100 maximum returns x 8 statistical methods x 900 parameter combinations). Selected results were then used in order to study the questions of interest. In particular we consider how the statistical methods compare in terms of eradicating data-mining bias, and furthermore study the effect of each simulation parameter on method performance. Firstly, it is worthwhile to observe how each simulation parameter affects the magnitude of data-mining bias in isolation.

4.1 Effect of Simulation Parameters on DMB

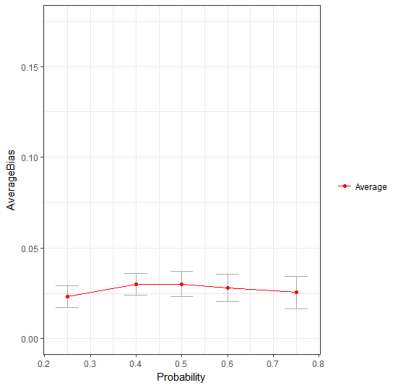
ATR return is determined purely by the ATR success rate. Since this rate was controlled at the outset, the expected return for each strategy is known and the degree of data-mining bias can thus be quantified. Magnitude of data-mining bias was calculated as the absolute value of $\bar{u_k}$ - $\mathbb{E}[\bar{u_k}]$, where $\bar{u_k}$ is the mean return produced by each ATR and $\mathbb{E}[\bar{u_k}]$ is given by,

$$\frac{1}{T}[p\sum_{t=1}^{T} u_{k,t} - (1-p)\sum_{t=1}^{T} u_{k,t}].$$

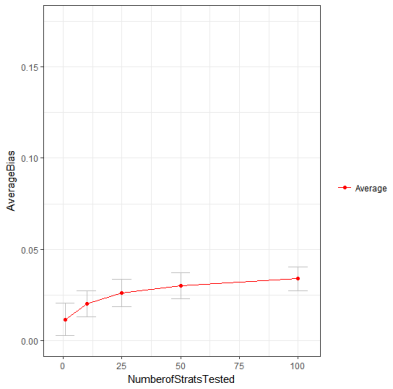
The figures below plot the relationship between each variable and the magnitude of data-mining bias present holding all other variables constant.

4.1.1 ATR Success Rate

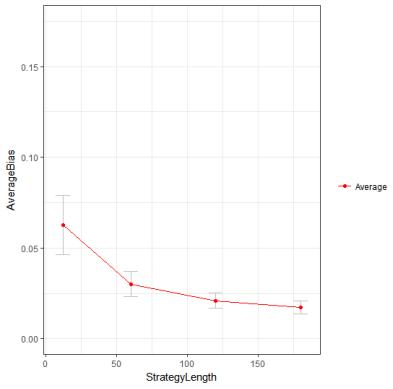
From Figure 1a it can be seen that the average magnitude of data-mining bias (DMB) is highest for an ATR success rate of 0.5. The DMB decreases on either side of this peak. The reason for this is that ATR success rate has



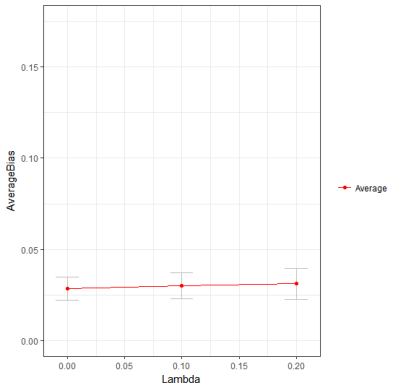
(a) Data-mining bias vs ART success rate



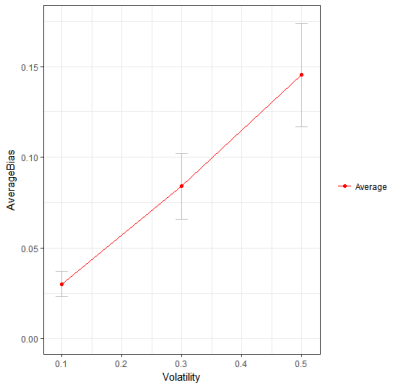
(b) Data-mining bias vs number of strategies



(c) Data-mining bias vs backtest length



(d) Data-mining bias vs presence of Outliers



(e) Data-mining bias vs volatility

Figure 1: DMB vs single parameter shifts

an indirect effect on DMB through its contribution to volatility. The closer ATR success is to 0.5, the higher the volatility of returns produced by the trading strategies. Figure 1*e* illustrates that volatility has a significant effect on average magnitude of DMB.

4.1.2 Number of Strategies Tested

Number of strategies tested refers to the value of K, from which a best-of-K strategy is selected. From Figure 1*b* it can be seen that, holding all else constant, the larger the number of strategies tested before a maximum is selected, the larger the magnitude of the DMB.

4.1.3 Length of Backtest

Varying the backtest length is analogous to varying the number of observations used to calculate the measure of performance used. As expected and as represented by Figure 1*c*, all else constant, the longer the backtest the smaller the magnitude of DMB. The shape of the curve also suggests that there may be a point at which the increased length of backtest may negate the effect of the other parameters and minimise DMB.

4.1.4 Presence of Outliers

The effect of an increase in the number of outliers was approximated by introducing jumps to the Merton Jump Model used to simulate market data. The frequency of these jumps was then increased (by manipulation of the λ parameter). All else constant the more frequent the jumps the larger the magnitude of the DMB, as shown in Figure 1*d*. However, this is a very small effect when compared to the other market parameters.

4.1.5 Volatility

The market volatility was changed by manipulating the standard deviation of diffusion (σ_d). From Figure 1*e* it can be seen that an increase in this parameter greatly increased the magnitude of DMB present in the results. Volatility is seen to have the largest effect on the magnitude of DMB and it will be worthwhile to see how the statistical methods perform in markets with high volatility.

Table 3: Significant rules identified by each method by parameter levels

(a) Base Case
$\{k = 50,\ T = 60,\ \lambda = 0.1,\ \sigma_d = 0.3\ \}$

p	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS
0.25	0	0	0	0	0	0	0	0
0.40	11	0	0	0	0	0	0	0
0.50	75	0	0	0	0	1	0	4
0.60	100	12	100	30	100	47	2	10
0.75	100	100	100	98	100	98	64	50

p	(b) $k = 25$								(c) $k = 100$							
	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS
0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.40	9	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
0.50	51	0	0	1	0	1	0	0	93	0	0	3	7	3	1	6
0.60	100	38	100	25	100	16	2	8	100	32	100	48	100	34	5	11
0.75	100	100	100	92	100	93	46	27	100	100	100	98	100	97	83	21
p	(d) $T = 12$								(e) $T = 180$							
	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS
0.25	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0.40	28	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0.50	55	0	0	0	0	3	3	4	87	1	1	1	65	1	0	1
0.60	100	38	97	2	2	4	3	2	100	92	100	100	100	95	78	33
0.75	92	9	66	2	100	6	1	2	100	100	100	100	100	100	100	86
p	(f) $\lambda = 0$								(g) $\lambda = 0.2$							
	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS
0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.40	3	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0
0.50	59	0	0	0	0	2	0	1	93	2	70	1	0	2	1	2
0.60	100	58	100	30	100	38	11	13	100	72	100	29	100	41	3	13
0.75	100	100	100	100	100	100	100	52	100	100	100	97	100	99	55	39
p	(h) $\sigma_d = 0.1$								(i) $\sigma_d = 0.5$							
	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS
0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.40	6	0	0	0	0	0	0	1	7	0	0	0	0	0	0	1
0.50	76	0	0	0	0	0	1	1	93	0	22	0	1	3	0	2
0.60	100	41	100	6	100	4	4	7	100	35	100	37	100	34	6	16
0.75	100	100	100	29	100	31	7	15	100	97	100	99	100	99	89	43

4.2 General Method Performance

Table 3 [8] shows method performance according to proportion of rules identified as significant for each ATR success rate p. Table 3a represents a base case where parameters were set as $\{k = 50,\ T = 120,\ \lambda = 0.08,\ \sigma_d = 0.1\}$. Tables 3$b$ - 3i show a change in each individual parameter holding all others constant in order to highlight the change in method performance.

In evaluating the methods, graphs representing the evolution of average p-value against a change in each of the parameters (all else held constant at the base values of $\{p = 0.5,\ k = 50,\ T = 60,\ \lambda = 0.1,\ \sigma_d = 0.3\ \}$) were also plotted. Figures $2a - 2e$ represent these relationships. FWER and FDR are methods that rely on altering bounds do not generate p-values and thus can not be included in these graphical representations.

For $p \leq 0.5$ we expect no strategies to be significant and can thus quantify the number of false discoveries made by each method. Table 4 records the number of false discoveries made by each method for each $p \leq 0.5$ as well as the total number (out of a total of 54 000 strategies) and proportion of false discoveries made. Table 4 includes expected false discoveries made due to the type 1 error rate[9] of $\alpha = 0.05$ and unexpected false discoveries made after accounting for α.

[8] The $p = 0.45$ and $p = 0.55$ cases were tested and produced results very similar to that of $p = 0.4$ and $p = 0.6$ respectively. As such, they have been omitted from the tables.

[9] The probability of making a false discovery as allowed for by the statistical methodologies.

Table 4: False discoveries (expected and unexpected) per method

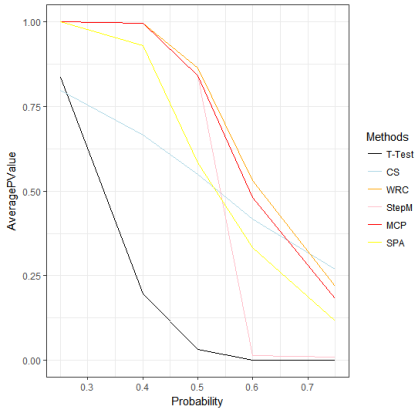
	Including expected								Unexpected							
p	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS	Ttest	FWER	FDR	WRC	StepM	MCP	SPA	CS
0.25	234	3	6	0	0	0	0	23	123	0	1	0	0	0	0	0
0.4	1979	101	398	1	0	6	13	87	1587	38	318	0	0	0	0	0
0.5	10601	662	3993	156	3178	220	78	371	9733	348	3602	9	2889	27	0	16
Total	12814	766	4397	157	3178	226	91	481	11443	386	3921	9	2889	27	0	16
Proportion	0.42	0.014	0.08	0.003	0.06	0.004	0.002	0.009	0.21	0.007	0.07	0.0001	0.05	0.0005	0	0.0003

4.2.1 t-test Performance

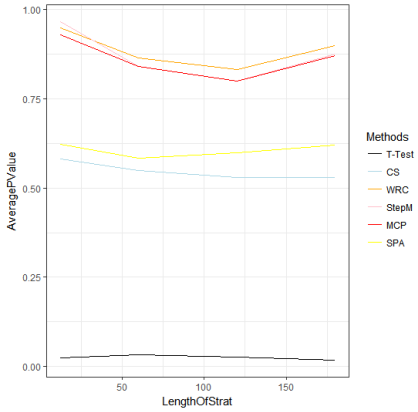
This method was implemented in order to show the effect of DMB on ordinary hypothesis tests. As can be seen in Table 4, for $p \leq 0.5$ (which suggest all strategies are unprofitable), the t-test is highly affected by DMB and 42% of strategies tested are identified as falsely significant. Our significance level is set at 0.05 and therefore we expect a maximum of 2700 false positives and the t-test produces 12814. The much larger number of false positives produced by the t-test give an indication of how susceptible these simple methods are to the effect of DMB. Furthermore, from Figures $2b - 2e$ we can see that the method does not improve according to changes in any simulation parameter values, as all average p-value curves move to 0, even though $p = 0.5$. From Table 3 it can be seen that the t-test has sufficient power for $p > 0.5$ as the strategies identified as significant are close to 100 in all sample cases however, this power comes with a massive penalty to accuracy when $p \leq 0.5$. Table 3 also displays that the test is highly affected by changes in parameters that increase DMB. All multiple hypothesis methods perform better than the t-test in terms of eradicating DMB and the results prove that the use of single hypothesis methods such as a t-test is not viable when data-mining.

4.2.2 Controlling the Family Wise Error Rate (FWER)

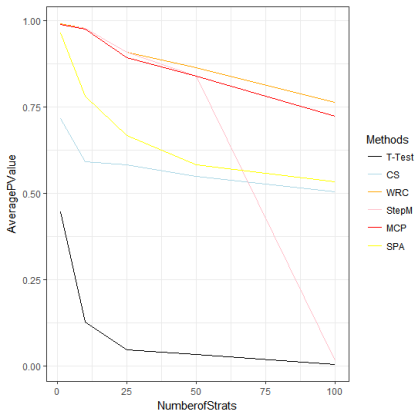
In testing 100 best-of-K strategies controlling the FWER using Bonferroni's correction has, in almost all cases, controlled for DMB. As shown in Table 3, in no instance in the specimen sample of parameter combinations has the test identified a greater than expected number of false discoveries for $p \leq 0.5$. However, FWER is shown to lack power especially in cases where $p = 0.6$, when all strategies should be identified as significant. Table 4 shows that there are cases for which controlling the FWER falls short as 766 false discoveries are made. Although this number corresponds to 1.4% of all strategies tested and is below the expected false discovery proportion of 5%, it is interesting to note that 386 (0.7%) unexpected false discoveries are made. Indicating that there are cases for which strategies produce far enough outlying returns to cross even the FWER adjusted critical value. This suggests that in these extreme cases bound modifying methods such as Bonferroni's correction may fall short. Methods that directly deliver appropriate p-values such as WRC and MCP may be more suitable in these situations.



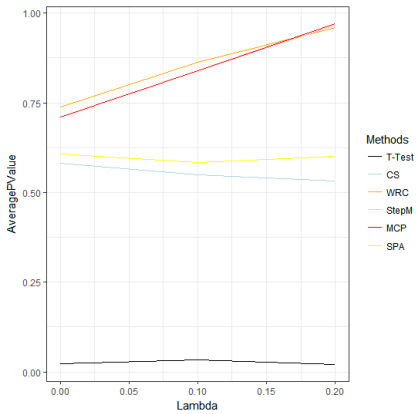
(a) Average p-value vs ATR success rate



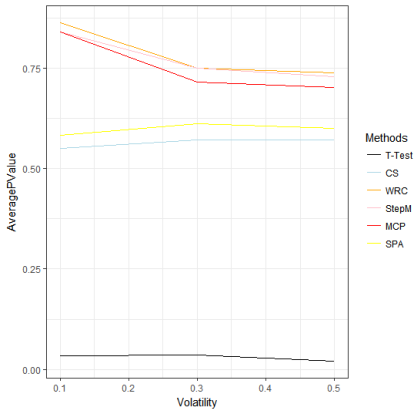
(b) Average p-value vs backtest length



(c) Average p-value vs number of strategies tested



(d) Average p-value vs outliers



(e) Average p-value vs volatility

Figure 2: Average p-values vs single parameter shifts

4.2.3 Controlling the False Discovery Rate (FDR)

Controlling the FDR was a proposed stepwise modification to controlling the FWER in order to improve the power of the test. Table 3 shows that FDR performs well for $p \leq 0.4$ and generally for $p \geq 0.6$ as almost perfect outputs are generated. However, in the case of $p = 0.5$ FDR is shown to be susceptible to the effects of changing simulation parameters that increase DMB (Table $3g$ and i). From Table 4 it can be seen that the increase in power of the test resulted in it producing 3921 unexpected false discoveries. This corresponds to just 7% of all strategies tested, so it may be seen as an appropriate sacrifice in accuracy in order to obtain the increased power. The results showed that under certain extreme stress tests FDR allows for too many additional false discoveries however, in most situations the test performs as expected and the increased number of false discoveries is offset by the power gained.

4.2.4 White's Reality Check (WRC)

White's Reality Check (WRC) is the first method employed for which p-values for each strategy are directly obtained from a sample distribution and bounds are not just modified. From Table 3 it can be seen that WRC performs well for $p \leq 0.5$ and a minimal number of false discoveries is produced. WRC identifies a low number of significant strategies on occasion when $p > 0.5$. This suggests a lack of power in certain conditions to identify possibly profitable strategies. Table 4 illustrates how well WRC eradicates DMB as it produces a proportion of 0.01% unexpected false discoveries. Figures $2b - 2e$ also show that WRC deals with changes in the simulation parameters well, as there are no drastic falls in the average p-value produced for changes to any parameter value. WRC is shown to be the least affected by DMB of all the methods for $p = 0.5$ as the average p-value curves produced are generally the closest to 1 over all parameter changes. WRC can be seen thus as a conservative method that can be used in any market condition in situations where accuracy is preferred over power.

4.2.5 Hansen's Test for Superior Predictive Ability (SPA)

Over the evaluation process Hansen's Test for Superior Predictive Ability (SPA) is shown to have very similar performance to WRC. From Table 4 it can be seen that SPA perfectly eradicates DMB as it makes 0 unexpected false discoveries. Hansen (2005) proposes that the SPA improves on the power of WRC. From Table 3 it can be seen that for $p \geq 0.6$ SPA identifies significant less profitable strategies than WRC on almost all occasions, contrary to Harris's claim. The reason for this is not exactly clear. This may be due to the fact that SPA proposes to improve on the power of WRC by disregarding the non-profitable strategies in the construction of the test statistic. However, in our best-of-K implementation there are no unprofitable strategies to be dis-

regarded and this may negatively impact the power of the test. Figures $2b-2e$ show that for $p = 0.5$ SPA does indeed have greater power than WRC however in these cases higher p-values are preferred and the increased power does not translate when $p \geq 0.6$. This suggests WRC and MCP are better performing methods under our implementation.

4.2.6 Monte Carlo Permutation (MCP)

Monte Carlo Permutation (MCP) produces results very similar to WRC over the evaluation process. The method also almost perfectly eradicates DMB as only 27 unexpected false discoveries are made (0.05% of all strategies tested). MCP's power is also very close WRC's power. This is shown in Table 3 from the very close number of significant strategies identified when $p \geq 0.6$. It is also shown in Figures $2a-2e$ as the average p-value curves produced by MCP generally lie very close to that of WRC. MCP can be seen as a conservative method that can be used in any market condition and is very similar to WRC, it may even be used to double check results when WRC is utilised as the methods are shown to almost perfectly correspond.

4.2.7 Romano & Wolf's Test (Step-M)

The Step-M method is a stepwise implementation of the WRC used in order to improve the power of WRC. Figure $2a$ shows that this power is indeed improved. Step-M produces perfect outputs in many of the specimen parameter combinations in Table 3, which indicates that in some cases this method will strike the perfect balance between power and accuracy. Due to the similar structure of the tests the Step-M method produces results very close to the results produced by FDR. When looking at Table 4 it is shown that the method also suffers the same pitfall. In extreme stress test situations the added power from the stepwise repetition of WRC has resulted in the method producing 2889 unexpected false discoveries (5% of all strategies tested). Figure $2c$ shows such a case as the method's accuracy is seriously impaired when $k = 100$ (p-value curve goes to 0). However, in all other cases the accuracy is maintained over changes in simulation parameters (p-value curve close to 1). This suggests that Step-M could be used in situations where WRC and MCP are considered too conservative.

4.2.8 Corradi & Swanson's Extension (CS)

Corradi & Swanson's Extension performs similarly to WRC. For $p \leq 0.5$ CS is shown to almost perfectly eradicate DMB in both Table 3 and Table 4. CS only produces 16 unexpected false discoveries out of 54 000. Corradi and Swanson (2011) propose that CS improves on the power of WRC. However, in Table 3 for $p \geq 0.6$, CS exclusively identifies a less than or equal number of significant strategies than WRC. This suggests that the method has less

power to identify significant strategies than WRC. Figures $2b - 2e$ show that the average p-value curve for CS is indeed below that of WRC as Corradi and Swanson (2011) suggest when $p = 0.5$, however this increased power does not seem to translate as p increases as displayed in Figure 2a (the p-value curve crosses above the curves for WRC and MCP when $p = 0.75$).

5 Conclusion

This study reviewed the various statistical methodologies in place to test multiple systematic trading strategies and implemented these methodologies under simulation with known artificial trading rules in order to critically compare and evaluate them. This evaluation was done using a best-of-K simulation method in order to simulate the real-world situation of investors selecting their best strategy N times. The results of the study showed that WRC and MCP were the best performing methods in terms of eradicating data-mining bias. Step-M was shown to be the most viable method for use in the case where WRC and MCP are too conservative. Claims made by Hansen (2005) and Corradi and Swanson (2011) that SPA and CS were more powerful than WRC were shown to not hold under our implementation. Furthermore, methods that rely on the manipulation of bounds such as FWER and FDR were shown to be inferior to their counter-parts that directly produce p-values such as WRC and Step-M in terms of eradicating DMB.

References

Aronson, D. (2011). *Evidence-based technical analysis: applying the scientific method and statistical inference to trading signals*, Vol. 274, John Wiley & Sons.

Bailey, D. H., Borwein, J. M., de Prado, M. L. and Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance, *Notices of the AMS* **61**(5): 458–471.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the royal statistical society. Series B (Methodological)* pp. 289–300.

Chordia, T., Goyal, A. and Saretto, A. (2017). p-hacking: Evidence from two million trading strategies.

Corradi, V. and Swanson, N. R. (2011). The white reality check and some of its recent extensions.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business & Economic Statistics* pp. 253–263.

Hansen, P. R. (2005). A test for superior predictive ability, *Journal of Business & Economic Statistics* **23**(4): 365–380.

Harris, M. (2016). Limitations of quantitative claims about trading strategy evaluation.

Harvey, C. R. and Liu, Y. (2013). Multiple testing in economics.

Harvey, C. R. and Liu, Y. (2014). Evaluating trading strategies, *The Journal of Portfolio Management* **40**(5): 108–118.

Masters, T. (2006). Monte carlo evaluation of trading systems.

Novy-Marx, R. (2016). Testing strategies based on multiple signals.

Peterson, B. (2015). Developing & backtesting systematic trading strategies.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping, *Econometrica* **73**(4): 1237–1282.

White, H. (2000). A reality check for data snooping, *Econometrica* **68**(5): 1097–1126.

A Appendix

A.1 Statistical Methods Additional Information

A.1.1 Controlling Family Wise Error Rate Illustrative Example

The key problem in multiple hypothesis testing is that the the probability of making a false discovery diverges from the initially defined α in line with the number of hypotheses tested.

For example, consider for a single hypothesis set $\alpha = 0.05$ and assume the null hypothesis is true. In this case we have that:

$$\begin{aligned} P(H_A|H_0) &= 1 - P(H_0|H_0) \\ &= 1 - 0.95 \\ &= 0.05 \\ &= \alpha. \end{aligned}$$

This differs from a multiple hypothesis test with $K = 10$. Again assume the null hypothesis is true in all cases, as well as the fact that all strategies are uncorrelated. The probability of making a false discovery is calculated in the following way,

$$\begin{aligned} P(H_A|H_0) &= 1 - \prod_{k=1}^{10} P(H_0|H_0) \\ &= 1 - \prod_{k=1}^{10} 0.95 \\ &= 0.4 \\ &>> \alpha. \end{aligned}$$

Intuitively in the context of systematic testing of systematic trading strategies this means that if 10 trading strategies are tested, there is a 40% chance that a strategy observed to be significantly better than the benchmark strategy is a false discovery.

The Family-Wise Error Rate (FWER) is defined as the probability of making at least one false discovery when testing multiple hypotheses (Harvey and Liu, 2013). In the above example it was shown that the FWER increased to 40% when 10 hypotheses were tested. Controlling the FWER is one of the fundamental tools needed to ensure that, when testing trading strategies, the number falsely identified to be significant can be bounded. The simplest way of controlling the FWER is known as the Bonferroni correction which states

that in order to preserve α as our type 1 error rate we should set $\alpha^* = \alpha/K$ and use α^* as our critical value (significance level) when performing each individual hypothesis test. For example, utilising the Bonferroni correction in the above example,

$$\begin{aligned} P(H_A|H_0) &= 1 - \prod_{k=1}^{10} P(H_0|H_0) \\ &= 1 - \prod_{k=1}^{10} (1 - \alpha^*) \\ &= 1 - \prod_{k=1}^{10} (1 - 0.05/10) \\ &= 0.05 \\ &= \alpha. \end{aligned}$$

This ensures that the FWER is preserved at α when testing multiple hypotheses. Incorporating the Bonferroni correction into a testing system for systematic trading strategies will thus limit the number of false discoveries to the pre-specified level α.

The problem with controlling the FWER in this manner is that it is very conservative. For example, if 1000 trading strategies were tested, in order to maintain an α of 0.05, α^* would need to be set at 0.00005. This indicates that a trading strategy would need to massively outperform the benchmark in order to reject the null hypothesis. Additionally, due to this level of conservativeness, a strategy that may be worth taking a risk on will also be excluded from the significant results. In our context allowing a few false discoveries into our pool of significant strategies in the hope of finding a greater number of profitable strategies is more viable than potential usage in the medical industry, where a single false positive drug, for example, could cause significant damage.

A.1.2 1st Order Stochastic Dominance

If, for any outcome y, there is at least as high a probability of achieving an outcome at least as good as y under A as under B (and if the probability is higher at some point), then A stochastically dominates B at the 1st order; formally:

$$F_A(y) \leq F_B(y) \forall y$$

and the strict inequality holds for some y.

A.1.3 Motivation Behind Controlling the False Discovery Rate

The False Discovery rate is defined as the expected proportion of falsely rejected hypotheses (Harvey and Liu, 2013). This is closely linked to the concept of statistical power, which refers to the probability that the statistical method will reject a false null hypothesis. The problem outlined above with controlling the FWER with the Bonferroni correction can also be stated in terms of power, in the sense that as more strategies are added the power of the test tends to zero. Controlling the FDR reduces the probability of false discoveries while not sacrificing as much of the statistical power of the hypothesis tests as controlling the FWER does. The principle way to control the FDR is by using the Benjamani & Hochberg Procedure (Benjamini and Hochberg, 1995). This procedure determines an upper bound for the number of false discoveries we have allowing us to increase the threshold level of rejection until that upper bound is as small as we require while still having the power needed.

A.1.4 Diebold & Mariano's Test

The Diebold & Mariano test was the initial framework introduced on how to test hypotheses regarding predictive ability. The framework is originally outlined by Diebold and Mariano (1995) in the context of forecast errors and can be altered in the following way to match our context. Consider testing two trading strategies against each other. Denote our variable of interest by $u_{k,t}$ for $t = 1, ..., T,$ and $k = 0, 1$ and the satisfaction associated with $u_{k,t}$ will be denoted $s(u_{k,t})$ and let $d_t = s(u_{1,t}) - s(u_{0,t})$ be the satisfaction differential. The null hypothesis implied by this test is,

$$H_0 : E(d_t) = 0,$$

with the alternative hypothesis being,

$$H_A : E(d_t) \neq 0.$$

Intuitively this this means that the null hypothesis is that the two trading strategies provide the same satisfaction and the alternative hypothesis is that the levels of satisfaction differ.

Diebold and Mariano (1995) state that H_0 should be tested according to the following test statistic,

$$\frac{\bar{d}}{\sqrt{\frac{2\pi f_d(0)}{T}}} \to N(0, 1),$$

where $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ and $f_d(0)$ is the spectral density of the satisfaction differential d_t at frequency 0 given by,

$$f_d(0) = \frac{1}{2\pi}(\sum_{j=-\infty}^{\infty} \gamma_d(k)).$$

The quantity, $\gamma_d(k)$ is the autocovariance of the satisfaction differential at lag k and is estimated by $\hat{\gamma}_d(k) = \frac{1}{T}\sum_{t=|k|+1}^{T}(d_t - \bar{d})(d_{t-|k|} - \bar{d})$.

This method was the initial building block upon which many of the following statistical methodologies such as WRC were built on. However, it is not applicable in our context of multiple hypotheses testing of systematic trading strategies. This is due to the method only being able to test 2 strategies against each other at a time, whereas we are concerned with testing $k > 2$ strategies. The Diebold & Mariano method also gives no indication of the directionality of the rejection of a hypothesis, simply that the satisfaction differs.