Avoiding data mining bias when testing technical analysis strategies - a methodological study

Rowan Douglas

Minor Dissertation submitted in partial fulfilment of the requirements for the degree of Master of Commerce

Section of Actuarial Science
University of Cape Town

September 2020

Abstract

When seeking to identify a profitable technical analysis (TA) strategy, a naïve investigation will compare a large number of possible strategies using the same set of historical market data. This process can give rise to a significant data mining bias, which can cause spurious results. There are various methods which account for this bias, with each one providing a different set of advantages and disadvantages. This dissertation compares three of these methods, the step wise Superior Predictive Ability (step-SPA) method of P.-H. Hsu, Y.-C. Hsu and Kuan (2010), the False Discovery Rate (FDR) method of Benjamini and Hochberg (1995) and the Monte Carlo Permutations (MCP) method of Masters (2006). The MCP method is also extended, using a step wise algorithm, to allow it to identify multiple profitable strategies. The results of the comparison show that while both the FDR and extended MCP methods can be useful under certain circumstances, the step-SPA method is ultimately the most robust, making it the best choice in spite of its significant computational requirements and stricter set of assumptions.

1 Introduction

Technical analysis (TA) has been used as an investment approach since the early twentieth century (P.-H. Hsu, Y.-C. Hsu and Kuan, 2010). In spite of this, the question of whether or not TA constitutes a profitable investment strategy is still up for debate, with evidence being found to support both sides of the argument. This, however, has not affected the use of TA in practice, with a large number of investors currently using some form of TA in their investment process (Menkhoff, 2010; Covel, 2004). These two facts mean that TA presents an interesting avenue for research, as the results will not only add evidence to the debate on TA strategies, but they will also be relevant to those who use TA in industry.

When conducting research on TA strategies, there are important considerations which need to be taken into account. The normal approach is to test a large number of different TA strategies, in an attempt to identify which, if any, are profitable. This approach makes use of statistical tests which are carried out on historical data. In the financial field, historical data generally consists of a single set of market or asset returns. This reliance on a single set of data means that the multiple hypotheses tested by the statistical test are dependent, which leads to the test producing spurious results. This situation and the negative effects it has on the validity of results is known as the data mining (DM) bias. Given how susceptible TA research is to the DM bias, it is imperative that this research takes the affects of the DM bias into account, in order to ensure the validity of any results found. There are multiple approaches which can be used in this regard, however, choosing one is not a straightforward task. Each approach provides a different balance between ease of implementation, the extent to which the DM bias is controlled, and how successfully profitable strategies are identified. This dissertation con-

siders some of the different approaches which can be used, to provide insight on how well they perform and to aid in making an informed choice between the different methods.

This dissertation considers four approaches: two are already-established methods, one is an established method which is extended by this dissertation, and the final method does not account for the DM bias, its inclusion being for comparative purposes. The first two methods are those of Benjamini and Hochberg (1995) and P.-H. Hsu, Y.-C. Hsu and Kuan (2010), both of which allow for multiple profitable strategies to be identified in a single investigation. The third method is that of Masters (2006). It only has the ability to identify a single profitable strategy in a single investigation. Therefore, this dissertation extends this method, using a step wise algorithm, enabling it to identify multiple profitable strategies in a single investigation. The final method is based on a standard student t-test. The performance of all four methods is compared using a simulation study, to allow for analysis of their relative strengths and weaknesses.

This dissertation makes the following two contributions: first, it extends the method of Masters (2006) to allow for the identification of multiple profitable strategies. Second, it provides a comprehensive comparison of the four methods considered, allowing for better informed decisions to be made between them for future TA research.

The dissertation proceeds as follows: first, TA and the DM bias are discussed briefly, before the standing literature on methods for investigating TA strategies is reviewed. Second, the methodology for implementing the four methods considered is discussed, including a description of the extension proposed by this dissertation. This is followed by a discussion of the methodology used for the simulation study. Finally, the results of the simu-

lation study are presented and the different conclusions which can be drawn from them are discussed.

2 Literature Review

2.1 Background on Technical Analysis

Aronson (2011, p. 1) defines technical analysis as "... the study of recurring patterns in financial market data with the intent of forecasting future price movements". This definition points to the simple process which stands at the core of TA. First, try to identify patterns in the price history of an asset. Second, see if any of these patterns are reliably followed by a specific price movement. Finally, if such a pattern can be identified, make trades based on the appearance of said pattern in live asset prices. The simplicity of this process, and the fact that there are an almost infinite number of methods for identifying patterns, has led to a multitude of different TA strategies being put forward by many different people.

Historically, most of these strategies fall into the category of what Aronson (2011) calls *subjective technical analysis*. These are TA strategies which do not have strict pattern definitions, but rather use the judgement of a practitioner to identify and interpret patterns. It is also not normal for subjective TA strategies to be accompanied by measurable claims regarding their performance. In most cases, the performance of a strategy is left open to interpretation and explanation. These two traits of subjective TA lead to two main problems. First, the fact that the identification of patterns relies on a practitioner, means that the strategies are not repeatable. This is because the same price history can be identified as displaying different patterns by different people, and sometimes even by the same person at different points

in time. Second, the fact that the measurement of performance for these strategies is often left open to interpretation means that it is difficult and often impossible to classify whether or not a certain strategy has worked. This problem is best illustrated with an example. Assume that a practitioner identifies pattern X, which they believe to be an indicator that an asset's price will rise. They will then buy an asset when pattern X manifests in its price history, stating that they believe that the price will increase for this asset. If in six months the asset's price has fallen, they can defend their strategy by claiming that the price increase has not had sufficient time to manifest, or by claiming that upon reflection the pattern showed different characteristics, which indicate a fall in price. The practitioner can do this because their initial claim for the rules performance, that the price would rise, is based on their own subjective interpretation and is not expressed in definite terms.

The two traits described above are problematic for subjective TA as they prevent any method of rigorous testing from being applied to the strategies. Therefore, one can never say definitively if a strategy is likely to work, and this, in turn, means that there is no useful way to select between subjective TA strategies.

To combat the shortcomings of subjective TA, Aronson (2011) puts forth the idea that practitioners need to move towards what he calls *objective technical analysis*. Objective TA strategies use clearly defined, repeatable methods for identifying patterns. Furthermore, objective TA strategies allow for performance to be accurately measured, without the results being open to interpretation. These two factors allow objective TA strategies to be rigorously tested using statistical methods, which in turn means that a definitive conclusion can be reached regarding whether or not an objective TA strategy

is likely to work. For an example of how objective TA can be approached, see the implementation of various TA strategies by Lo, Mamaysky and Wang (2000). While using objective TA is advantageous due to the definitive conclusions which can be reached, it does introduce a new problem: the need to ensure that the statistical tests used are set up to be accurate and free of bias. It is this problem which leads to the methods being investigated in this dissertation. In the interest of brevity, for the remainder of this dissertation, *TA strategies* will be used to reference objective TA strategies, unless specifically stated otherwise.

2.2 The Data Mining Bias

One of the main contributors to the problem of inaccurate, biased statistical tests in TA investigations is the DM bias discussed in the introduction. Recall that the DM bias refers to the negative effect which testing multiple dependent hypotheses can have on the results of a statistical test. To understand what these negative effects are, it is necessary to briefly describe the procedure for a standard statistical test.

A standard statistical test has two hypotheses, a null hypothesis and an alternative hypothesis, with only one of these hypotheses accurately describing reality. That is to say that, in reality, only one of the two hypotheses is true. The aim of the test is to identify which of the two hypotheses is supported by the evidence, with the hope of correctly identifying the true hypothesis. If a test identifies that the null hypothesis is supported by the evidence, it is said to accept the null hypothesis. If, however, a test identifies that the alternative hypothesis is supported by the evidence, it is said to reject the null hypothesis. These two possible results for a test, combined with the fact that in reality the null hypothesis is either true or false, leads

to four possible outcomes for a test. These outcomes are illustrated in table 1, labelled alphabetically from A to D. Outcome A occurs when the test accepts the null hypothesis and in reality the null hypothesis is true. Outcome B occurs when the test accepts the null hypothesis and in reality the null hypothesis is false. Outcome C occurs when the test rejects the null hypothesis and in reality the null hypothesis is true. Outcome D occurs when the test rejects the null hypothesis and in reality the null hypothesis is false. Of these four outcomes, A and D result in a correct conclusion, while B and C result in an incorrect conclusion. Outcomes B and C, therefore, can be seen as errors, with outcome C referred to as a type I error, and outcome B referred to as a type II error.

The rates of type I and type II errors are not independent, as reducing the level of one causes an increase in the level of the other. Therefore, it is necessary to prioritise controlling one of these errors. Due to the fact that incorrectly rejecting the null hypothesis is judged as being the more severe error, statistical tests prioritise controlling the type I error. They accomplish this by first requiring a level, α, to be chosen. The test then uses this level to determines whether or not the evidence considered is sufficient to reject the null hypothesis and state that the alternative hypothesis is true. What is considered as sufficient evidence is determined by the value of α, with the test only rejecting the null hypothesis if the evidence shows that this would be an accurate result $(1-\alpha)\%$ of the time. This ensures that the type I error is controlled at a level less than or equal to α. This, in turn, means that conclusions can be drawn with a set level of confidence that a type I error has not been made.

What happens under the effects of the DM bias, however, is that even though a test attempts to control the type I error at the level of α, it fails

Table 1: Illustration of the four possible hypothesis test outcomes.

	Null True	Null False
Null Accepted	A: Correct	B: Type II error
Null rejected	C: Type I error	D: Correct

to do so, and the actual type I error is larger than α. A test which has this problem is said to be anti-conservative. Any conclusions which are drawn from the results of an anti-conservative test are invalid, as they are based on a level of confidence which is incorrect. In practice, this means that the danger of the DM bias for TA research is that it can lead to an investigation identifying strategies as being profitable, when in reality they are not. This danger is illustrated by Harvey and Liu (2014), who provide examples of some misleading conclusions which can be found if the DM bias is not accounted for. A further illustration of this danger is presented in the findings of Chordia, Goyal and Saretto (2017). They tested over two million trading rules and found that many of those which at first appeared to provide significant returns, were no longer found to be significant once methods which account for the DM bias were used.

While the above discussion used a test of a single hypothesis for illustrative purposes, in reality, investigations which are susceptible to the DM bias test multiple hypotheses. In order to discuss the frequency of errors for tests which consider multiple hypotheses, it is necessary to define certain rates. The first of these rates is the *average power* of a method. This rate is defined by Romano and Wolf (2005) as the expected proportion of the true alternative hypotheses which the method correctly accepts. This rate gives the inverse of the type II error for a test of multiple hypotheses. In

the interest of brevity, *average power* will be referred to simply as power for the remainder of this dissertation. To measure power in the simulation study carried out by this dissertation, the Average Rejection (AR) rate is used. The AR rate is the average, across all replications of a simulation, of the number of true alternate hypotheses accepted as a proportion of all the true alternate hypotheses considered. That is to say, the AR rate gives the average power achieved by a method in a simulation study.

The next two rates which need to be defined both refer to how likely it is that a true null hypothesis will be rejected. Therefore, they are both linked to the type I error for a test of multiple hypotheses. The first of these rates is the Family Wise Error (FWE) rate, which is defined as the proportion of times which the test would reject at least one true null hypothesis if it were repeated multiple times. Therefore, the FWE rate indicates the likelihood that the test will incorrectly reject at least one of the null hypotheses considered, effectively measuring the type I error for a test of multiple hypotheses. The second of these rates is the False Discovery (FD) rate, which is defined as the proportion of the total number of rejected null hypotheses which were true null hypotheses. Therefore, the FD rate indicates the likelihood that each of the rejections made by a method were made incorrectly. As such, controlling the FD rate for a test of multiple hypotheses does not control the type I error.

An example can be used to illustrate the difference between the FWE and the FD rates. While in practice the rates describe expectations, this example expresses them as exact numbers across many replications in order to simplify the explanation. Consider an investigation which conducts a multiple hypothesis test with 10 hypotheses. Then imagine that this investigation is replicated 100 times. This results in a total of 1 000 hypotheses across all

the replications. If the investigation controls the FWE rate at a level of 5%, this means that there will be only 5 replications out of 100 where a type I error occurs. In the remaining 95 replications, there will be no type I errors. Alternatively, if the investigation controls the FD rate at a level of 5%, this means that 50 out of the 1 000 hypotheses tested will have a conclusion which is a type I error. In the worst case scenario, these 50 hypotheses could be spread evenly across the 100 replications, resulting in 50 replications in which one type I error occurs, and only 50 remaining replications without a type I error. Therefore, in this example, the difference between controlling the FWE rate and the FD rate is the difference between ensuring that 95 out of 100 replications have no type I errors compared to ensuring that 50 out of 100 replications have no type I errors. This shows that controlling the FD rate is a less strict threshold than controlling the FWE rate. This weaker threshold has the benefit of allowing for a more powerful investigation method. However, it does mean that the results of the method have to be interpreted in a way which accounts for the fact that it is the FD rate which was controlled and not the FWE rate. This is an important point, as the convention for interpreting the results of statistical tests is based on the assumption that the FWE rate is the rate which was controlled.

A final point to make when discussing the DM bias is the use of the terms data snooping and data mining. Both terms are used in different contexts to refer to the idea described as DM bias in this dissertation. P.-H. Hsu, Y.-C. Hsu and Kuan (2010) point out that the term data mining has recently taken on an alternative meaning in the context of big data, and advocate for the use of data snooping instead. However, in some cases, the term data snooping has been used to refer to the act of using the results from previous studies to inform the choice of strategy to be included in an investigation. Therefore,

this dissertation chooses to use the term data mining over the term data snooping.

2.3 Methods for Investigating TA Strategies

The susceptibility of TA strategies to the DM bias necessitates the use of investigation methods which take the DM bias into account. As was discussed in section 2.2, the DM bias causes true null hypotheses to be rejected too often, resulting in a higher than expected type I error rate. Therefore, methods which account for the DM bias control how often this occurs. In this dissertation, this practice will be referred to as controlling for the DM bias, however, it could also be referred to as controlling the type I error. Controlling the results in this way comes at a cost, however, as it decreases the likelihood that a true alternative hypothesis will be identified. Harvey and Liu (2014) point out that this is the typical type I error versus type II error trade off. This trade off was mentioned in section 2.2, where decreasing one of the errors results in an increase in the other. As a consequence of this trade off, investigation methods cannot simply aim to control for the DM bias. Instead, it is necessary that they seek to strike a balance between controlling for the DM bias and ensuring that they achieve an acceptable level of power. In the ensuing discussion of various investigation methods, their ability to successfully strike this balance is the primary factor on which they are judged.

Note that due to the fact that the causes of the DM bias are not unique to investigations of TA strategies, most of the methods discussed below focus on controlling the results of investigations which consider multiple hypotheses, a broader category which investigations of TA strategies fall into.

2.4 Categories of Investigation Methods

When discussing methods for investigating TA strategies, it is useful to split the methods into different categories. P.-H. Hsu and Kuan (2005) point out that the methods which account for the DM bias when testing multiple hypotheses can be split into two main categories based on whether they alter the data or the methodology of the test. The first category involves designing the study in such a way that the different tests do not rely on the same set of data. This can be achieved by using different, but comparable data sets for each test.

For example Lakonishok, Shleifer and Vishny (1994) carried out their study using data from both the New York stock exchange and the American stock exchange. This, however, requires comparable data sets, which are not always available, especially for TA studies. If comparable data sets are not available, a second option is to split a single data set into subsets, and compare the performance of the strategies across the subsets to verify the results. For example Fernández-Rodriguez, González-Martel and Sosvilla-Rivero (2000) compare their results across three subsets and adjust their conclusions accordingly. This second option is also not ideal for TA as the choice of how to split the data is arbitrary, which can affect the objectivity of the tests. This is noted by both Aronson (2011) and P.-H. Hsu and Kuan (2005). A third and final option for adjusting the data set is to use one subset of data to choose the best performing TA strategies and a second separate subset to measure their unbiased performance. As pointed out by Aronson (2011) this option also has the drawback of arbitrarily splitting the data, a choice which can materially affect the results of the investigation. He also points out a further drawback: the fact that the use of a testing subset means that the most recent data cannot be used to inform the choice of strategy,

as it is only used for testing. The different drawbacks for all three of these options mean that none of them are ideal for investigating TA strategies, therefore, this dissertation focuses on the methods from the second category, described below.

The second category of methods use all of the available data, instead adjusting the testing methodology in order to control for the DM bias. The methods in this category can be further split into adjustment methods, which alter the α value against which p-values are tested, and randomisation methods, which use randomisation techniques to adjust the null distribution for the hypothesis tests. The methods which will be covered in this dissertation come from these two subcategories, as such, each subcategory is discussed in a separate section below.

2.4.1 Adjustment Methods

The most widely known method in this category is the Bonferroni adjustment. For a test of m strategies, the Bonferroni adjustment states that a critical level of $\alpha' = \frac{\alpha}{m}$ should be used. While this adjustment does control for the effects of the DM bias, it also results in a very low power for the test. Both White (2000) and P.-H. Hsu and Kuan (2005) note that this makes the Bonferroni adjustment inappropriate for investigations of TA strategies, due to the large number of strategies which are normally considered.

A more powerful adjustment method, labelled the BH method for convenience, is suggested by Benjamini and Hochberg (1995). They note that the Bonferroni adjustment controls the FWE rate. Therefore, they suggest that the FD rate be controlled instead, to allow for a more powerful method. In their paper, Benjamini and Hochberg (1995) acknowledge that controlling the FD rate will result in a high type I error. However, they argue that there

are certain cases where this is an acceptable compromise. These cases are situations where the conclusion of the investigation does not rely on each of the hypotheses which are rejected being true alternate hypotheses. In these situations, a possible incorrectly rejected null hypothesis will not invalidate the entire investigation. Therefore, Benjamini and Hochberg (1995) argue that the increased type I error is worth the additional power which is gained.

Investigations of TA strategies can fall on either side of this classification depending on how the results of the method are used. If only a single strategy from those identified as being significant in the investigation will be used, then the compromise inherent in the BH method would not be acceptable. If, however, as many as possible of the strategies identified as being significant will be used, then the fact that a small portion of the strategies might not be truly significant is not entirely detrimental and it could be argued that the compromise inherent in the BH method is acceptable. Due to the fact that the increased power of the BH method is required to observe meaningful results when testing strategies on the scale that is normal for a TA investigation, it is the adjustment method chosen to be analysed in this dissertation, with the caveat that if it is used in practice, the results of the method should be used in the appropriate manner as described above. The specifics of the BH method are discussed in more detail in the methodology section of this dissertation.

2.4.2 Randomisation Methods

The first test in this category is also the first test to be discussed which was designed with financial predictive models as its primary focus. As noted by Hansen (2005) the term *model* is used in this context as a broad term covering all types of financial rules and methods. As such, the method can be easily

applied to investigations of TA strategies. This test is known as White's Reality Check (WRC) and it was put forward by White (2000). White noted that adjustment methods do not perform well given the number of comparisons usually required in financial investigations. As such, he sought to develop a method which directly produced appropriate p-values, thus negating the need for adjustments to be made. He achieved this by setting up the test in a way which allows for the null distribution to be estimated using randomisation techniques. In particular, his preferred method uses bootstrapping to attain an empirical null distribution. The null hypothesis for the WRC is that the performance of the best model, out of all those considered, is no better than that of a benchmark. The performance of a model (or rule in the case of a TA investigation) can be defined using any metric which can be calculated from the model's time series. While the WRC method does provide a better test than the adjustment methods, it does have two drawbacks which limit its performance. These drawbacks and the alternative methods which address them are discussed below.

Hansen (2005) points out that the formulation of the WRC method results in an unnecessary loss of power. This is due to the fact that the WRC method constructs its null distribution under the assumption that all models have performance exactly equal to that of the benchmark. This assumption does not allow for the fact that some models may underperform the benchmark. Constructing the null distribution in this way is known as constructing it using the Least Favourable Configuration (LFC). Using the LFC means that even if models perform much worse than the benchmark, they are taken as being as good as the benchmark when constructing the null distribution. This causes the null distribution to be overly conservative. Hansen (2005) addresses this by suggesting the Superior Predictive Ability (SPA) method,

which makes two adjustments to the WRC method. First, the SPA method uses a sample-dependent null distribution which reduces the influence on the null distribution of models which the data suggests are significantly worse than the benchmark. Second, the SPA method uses a studentised test statistic. Hansen (2005) uses both a simulation and an empirical study to show that the SPA method does indeed provide a more powerful test, while still controlling for the effects of the DM bias.

The second drawback of the WRC method is raised by Romano and Wolf (2005). They point out that the WRC method only allows for the identification of a single outperforming model. To address this drawback they enhance the WRC method using a step wise algorithm which allows for multiple outperforming models to be identified. This method is labelled as the step-RC method. To further improve the power of their step-RC method, Romano and Wolf (2005) also suggest the use of a studentised test statistic, which they implement in their test. Romano and Wolf (2005) show that the step-RC method does indeed have better power compared to the WRC method, while still controlling for the effects of the DM bias.

While the SPA and step-RC methods do address the drawbacks of the WRC method in isolation, neither method accounts for them both. This led to the step-SPA method of P.-H. Hsu, Y.-C. Hsu and Kuan (2010), which combines the improvements made in the SPA method with the step wise algorithm of the step-RC method. P.-H. Hsu, Y.-C. Hsu and Kuan (2010) show that the step-SPA method is more powerful than both the SPA method and the step-RC method, while still controlling for the effects of the DM bias. This suggests that the step-SPA method is a better choice than either the SPA method or the step-RC method when attempting to address the drawbacks of the WRC method. While this cannot be taken to mean that

the step-SPA method is the best possible approach based on that of White (2000), it does indicate that the step-SPA method is the best among those considered for this dissertation. Further methods based on this approach were not considered in order to control the scope of the dissertation. As such, the step-SPA method is used to represent randomisation approaches based on the approach of White (2000) in the analysis carried out in this dissertation. The specifics of the step-SPA method are discussed in more detail in the methodology section of this dissertation.

An alternative testing method which is also based on randomisation techniques, but which does not stem from the WRC method, is the Monte Carlo Permutation (MCP) method used in the book *Evidence-based Technical Analysis* written by Aronson (2011). The MCP method appears to have been developed specifically for Aronson's book with the aid of Dr. Timothy Masters. In a document which is published on the book's website, Masters discusses the MCP method in far greater detail than Aronson does in the book itself. Therefore, this document (see Masters, 2006) is used in this dissertation as the reference for the MCP method. The MCP method focuses exclusively on trading strategies, considering the trading signals they generate for each period and the possible return to be earned in each period. Using this information, the MCP method tests the null hypothesis that all of the strategies considered are in fact random strategies, i.e. strategies where the trading signals have been chosen randomly. The alternate hypothesis is that the best performing strategy from the set being considered is not a random strategy. The motivation for these hypotheses is based on the idea that in order for a strategy to be profitable, it must make informed trading decisions, with the opposite of informed decisions being random decisions.

In order to test this null hypothesis, the MCP method approximates the

distribution for the return which would be earned by the best performing strategy, given that all the strategies considered are random strategies. This distribution is the null distribution. The return for the best performing strategy from those considered is then compared against the null distribution to calculate the probability of observing such a return if the null hypothesis is true. The process for approximating the null distribution is now described in more detail as the assumptions which underlie the process have an affect on the performance of the MCP method.

The process approximates the distribution by finding a large number of the possible values which belong to it. Each of these values is found by repeating the following steps. First, the trading signals from each of the rules are randomly paired with a period return. Second, the strategy returns resulting from these random pairings are calculated, giving a set of random strategy returns. Finally, the best return from this set of random strategy returns is recorded and used as the distribution value for this iteration. This process relies on the assumption that the pairings in step one are all possible in practice. If they are not, then the random strategy returns found using these pairs are not an accurate representation of the random strategies which could occur in reality, and the distribution produced will not be a good approximation of the true null distribution. Some of the situations which can violate this assumption are discussed below when covering the requirements of the MCP method.

When considering why the MCP test is a good alternative, Masters notes that methods which rely on bootstrapping have an inherent weakness in the fact that they assume that the distribution of the sample is representative of the overall population distribution. Methods which use a Monte Carlo permutation have a much smaller reliance on the sample distribution, giv-

ing them an advantage over bootstrapping methods in this regard. This advantage does come at a cost, however, as the MCP method requires the models being tested to meet some fairly strict requirements. First, the models must represent some form of trading strategy which produces a signal for each trading period. These signals are usually classified as one of buy, sell or neutral, but the only requirement is that there are at least two different kinds of signal. Second, a complete history of the strategies signals for each period as well as the ability to calculate potential profit or loss for each period is required. Third, for any period, the trading signal chosen by a strategy should not be limited by whether or not the strategy already initiated a position in a previous period. This requirement is important as it prevents the assumption mentioned in the previous paragraph from being violated. The first and second requirements are not particularly difficult to meet for investigations of TA strategies. The third requirement is not as easily met. However, in general it does not pose a problem for investigations of TA strategies, as long as the strategies are set-up with the requirement in mind. This involves formulating the strategies in such a way so that the trading periods cannot overlap, ensuring that the signals chosen are independent of the previous positions taken by the strategy. This can be achieved by defining a fixed holding period for each position, and only generating a new signal once this fixed period has expired.

Due to the fact that the MCP method is less well known than the WRC method, its potential drawbacks have not been investigated as widely. However, as with the WRC method, the MCP method only has the potential to identify a single outperforming strategy. Therefore, the power of the MCP method could also be improved by altering the method to allow for the identification of multiple outperforming strategies. Romano and Wolf (2005) note

that the step wise algorithm which they apply to the WRC can apply to other methods which consider multiple models or strategies. This dissertation will extend the MCP method to include an adaption of the step wise algorithm resulting in the step-MCP method. The details of this extension are covered in the methodology section.

There are also two further potential drawbacks of the MCP method. First, it is not clear how the power of the MCP method will be affected if poor strategies are present in the set being considered. If the MCP method has the same drawback as the WRC method, causing it to lose power when poor strategies are present, then evidence of this should be seen in the simulation study performed in this dissertation. The second further potential drawback of the MCP method is raised by Masters (2006) himself. He notes that if the returns on which the TA strategies are based display autocorrelation, then there is the chance that the MCP method could become anti-conservative. That is to say that it will not correctly control for the DM bias. This happens because, even if all of the strategies considered are indeed random strategies, some of them may have the same trading signal for multiple consecutive periods purely by chance. When these runs of signals happen to align with a run of appropriately signed returns, which arise due to the autocorrelation, then the strategy will have an abnormally high return, even though it is a random strategy. This behaviour is lost when constructing the null distribution, however, as the random pairing does not take any autocorrelation of returns into account. Therefore, the random strategies which were lucky enough to match a run of trade signals with a run of returns will appear to be significant more often than they should, resulting in an anti-conservative test. In order to check the extent of this drawback, the performance of the step-MCP method, along with the other methods considered, is tested using

varying degrees of autocorrelated returns in the simulation study.

2.5 Comparisons of TA Investigation Methods

The majority of existing literature which compares methods for investigating TA does so in order to demonstrate the performance of a new method which is detailed in said literature. A single example was found of a paper which carried out a comparison purely for the sake of assessing already existing methods (Perumal and Flint, 2018).

Regardless of the aims of the comparison, the most common approach used to compare the performance of different methods for investigating TA strategies is to carry out a simulation, where the true nature of the strategies being investigated is known in advance. This allows for the results of the various methods to be judged based on how many of the strategies they correctly classify as being either profitable or not profitable. There are a number of parameters which are required for these simulations, such as the methods which will be compared, the number of simulations to be run, the number of rules which will be simulated, the length of the data series generated for each rule and how the data required is generated. The values chosen for these parameters can have a material impact on the results of the simulation, therefore, it is important to consider what the values chosen are, and why these choices were made. The impact of these parameters can be identified by running multiple simulations, each with different parameter values.

Examples of papers which use a simulation to demonstrate performance of a new method are Romano and Wolf (2005), Hansen (2005) and P.-H. Hsu, Y.-C. Hsu and Kuan (2010). The new methods assessed in each of these papers have already been discussed in the previous section. While each of the

papers does use a different set of simulation parameters, the overall structure of the different simulations is very similar. There is a set number of simulated rules which are tested, the returns for which are generated randomly using a set algorithm. The various investigation methods being compared are then carried out on these generated rule returns. Certain metrics for each of the methods, such as the FWE rate, are calculated, and these are used to draw conclusions regarding the comparative performance of the methods. The specific conclusions for these simulations have already been noted in the discussion of the respective papers above.

The above mentioned paper, which focuses solely on comparison of methods, is that of Perumal and Flint (2018). Their simulation follows a modified approach due to the fact that they include the MCP method in their analysis. This requires a modified approach due to the fact that WRC-based methods only require the generation of the period returns for each trading rule, while MCP-based methods require the generation of both an overall asset return and a series of rule signals for each rule. Therefore, Perumal and Flint (2018) split the return generation process for the simulation into two steps. First, they generate the asset returns, using a similar method to how the rule returns are generated for WRC-based simulations. Second, they generate a series of buy or sell signals for each of the rules being simulated. These rule signals are applied to the asset returns generated in the first step to calculate the final return for each of the rules. This provides the necessary data for both the WRC and MCP-based methods to be simulated.

While the modified approach of Perumal and Flint (2018) is appropriate for the rules being tested, their simulation was designed to test systemic trading strategies, not TA strategies. This means that some of the generation methods used and the parameter values chosen for the simulation are

not appropriate for a study of TA strategies. These elements provide scope for this dissertation to expand upon the investigation of Perumal and Flint (2018) by changing them to be in line with the context of this dissertation.

The first of these elements is the fact that the rules generated for the simulation are not an accurate representation of what would be expected when using a TA strategy. This is due to two factors of the simulation process; the method for generating rule signals and the choice of rules to include in each simulation. The method for generating rule signals is problematic due to the fact that it makes the assumption that if a rule does not produce the correct signal, then it must produce an incorrect signal. This does not allow for the case where the rule outputs a profitable signal not because it made a correct choice, but rather due to luck. This is a common occurrence with TA strategies, therefore, it is important to incorporate this characteristic in simulated TA rules. Another effect of this assumption is that it does not allow for the generation of a neutral signal. Many TA strategies do output neutral signals, therefore, it is once again important to incorporate this characteristic in simulated TA rules. A better assumption to make, in order to generate rules which are appropriate for the TA context, is to assume that if a rule does not generate the correct signal, then it may produce a buy, sell or neutral signal with equal probability.

The choice of rules to include in each simulation is problematic as only a single type of rule is included in each simulation. That is to say that either only unprofitable, only zero-profit or only profitable rules are included. This is a problem in the context of TA investigations as once again, it is not in line with what is expected to occur in practice. When testing a set of TA strategies in reality, there will be a large number of rules, most of which will be zero-profit rules, with the potential for some profitable and some

unprofitable rules as well. This is seen in multiple real-world tests, including those carried out by Hansen (2005), P.-H. Hsu, Y.-C. Hsu and Kuan (2010) and Aronson (2011).

The other elements from the investigation of Perumal and Flint (2018) which are not appropriate in the context of TA strategies are related to the values chosen for certain parameters. While none of these elements present a major issue, they are still areas which can be improved upon. Each of these is discussed briefly below. First, the number of periods considered for the simulation is relatively small, with the maximum number of periods for any simulation being 180, chosen as 15 years of monthly data. This is not an ideal value, however, due to the fact that most TA strategies are carried out on daily data, therefore, most tests will consider a much larger number of periods. Second, the number of replications in each simulation is 100, which is relatively low. This is not ideal as it means that the results of the simulation are not as robust as they could be. Finally, the number of rules considered for each simulation is also relatively small, with the maximum number for any simulation being 100. This is not ideal as TA investigations can often consider thousands of rules. While computational limits may not allow for thousands of rules to be considered for each simulation, a larger number would provide a better representation of reality.

Given that Perumal and Flint (2018) set the above elements to be appropriate in the context of systematic trading strategies, rather than TA strategies, it is not surprising that the results they found are not entirely in line with those from papers which focused on the context of TA strategies. For example, while they do find that the WRC and MCP methods perform similarly, they also find that both the SPA and Step-WRC methods do not outperform the WRC method. While these results are unexpected given the

results from the other literature considered, this dissertation will not investigate these discrepancies for the following reasons. First, both the SPA and step-WRC methods are shown to outperform the WRC method in their respective papers (Hansen, 2005; Romano and Wolf, 2005). Second, Perumal and Flint (2018) acknowledge that their results may be affected by the implementation of their test. Finally, investigating further WRC-based methods is outside the scope of this dissertation (as discussed in section 2.4.2).

3 Methodology

3.1 Implementation of Investigation Methods

This section covers the implementation of the four investigation methods considered for this dissertation.

3.1.1 t-test

The t-test can be used on each individual strategy included in an investigation. Each strategy is then classified as being profitable or not based on whether a one-sided t-test finds the mean return for the strategy to be significantly greater than zero. This method of investigating TA strategies does not account for the DM bias. As such, it is included in the analysis purely as a baseline, to illustrate the severity of the problems encountered if the DM bias is not accounted for. For the remainder of this dissertation, this method will be labelled as the T method.

3.1.2 BH Method

The BH method was introduced in section 2.4.1 of this dissertation. It adjusts the results of individual hypothesis tests according to the following process.

First, an individual test is carried out on each strategy. For this dissertation, the T method described in the previous section is used to conduct the individual tests. Once these individual tests have been carried out, their respective p-values are ordered from smallest to largest. Following the notation used by Benjamini and Hochberg (1995), let $P_{(1)} \leq P_{(1)} \leq \ldots \leq P_{(m)}$ denote the ordered p-values for a test of m strategies, with $H_{(i)}$ denoting the hypothesis which produced $P_{(i)}$. A decision on which of the hypotheses can be rejected is then made by finding the value of k such that it is the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha$, where α is the critical value for the test. The hypotheses $H_{(i)}$ $\forall$ $i = 1, 2, \ldots, k$ can then be rejected.

3.1.3 Step-SPA Method

The step-SPA method was introduced in section 2.4.2, and its implementation for this dissertation is detailed below. The notation used in this description follows the notation used by P.-H. Hsu, Y.-C. Hsu and Kuan (2010). First define the following:

$$
\begin{aligned}
m &: \text{ Number of strategies being tested.} \\
k &: \text{ Subscript for the strategy number, } k = 1, 2, \ldots, m. \\
n &: \text{ Number of periods for each strategy.} \\
t &: \text{ Subscript for the period number, } t = 1, 2, \ldots, n. \\
d_{k,t} &: \text{ Return for strategy } k \text{ during period } t. \\
\mu_k &: \mathbb{E}(d_{k,t}) \text{ for all } t. \\
\boldsymbol{d}_t &: (d_{1,t}, d_{2,t}, \ldots, d_{m,t})'
\end{aligned}
$$

Note that P.-H. Hsu, Y.-C. Hsu and Kuan (2010) define $d_{k,t}$ as a performance measure relative to a benchmark, however, for the purposes of this disserta-

tion a benchmark of zero is used, so the performance measure is simply the strategy's return.

Before continuing with the definitions, it is necessary to detail an assumption which is required for the step-SPA method. The necessity of this assumption is noted by P.-H. Hsu, Y.-C. Hsu and Kuan (2010) and it can be described as follows: assume that $\boldsymbol{d}_t$ is strictly stationary and α-mixing of size $\frac{-(2+\eta)(r+\eta)}{(r-2)}$, for some $r > 2$ and $\eta > 0$, where $E|\boldsymbol{d}_t|^{(r+\eta)} < \infty$ with $|\cdot|$ the Euclidean norm, and $\text{var}(d_{k,t}) > 0$ for all k. This assumption allows $\boldsymbol{d}_t$ to exhibit weak dependence over time, and is necessary for the following two reasons. First, under this assumption, the following is true as a result of the central limit theorem:

$$\sqrt{n}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}) \xrightarrow{\text{D}} N(\mathbf{0}, \boldsymbol{\Omega})$$

where $\bar{\boldsymbol{d}} = \frac{\sum_{t=1}^{n} \boldsymbol{d}_t}{n}$, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{d}_t)$, $\boldsymbol{\Omega} \equiv \lim_{n\to\infty} \text{var}(\sqrt{n}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}))$, and $\xrightarrow{\text{D}}$ represents convergence in distribution. Second, this assumption allows for the use of both the stationary bootstrapping procedure and the covariance matrix estimator of Politis and Romano (1994), which are used in the step-SPA method.

Using the consequences of this assumption, the following two further definitions can be made:

$$\begin{aligned} \widehat{\boldsymbol{\Omega}} &: \quad \text{A consistent estimator for } \boldsymbol{\Omega}. \\ \hat{\omega}_{ij} &: \quad \text{The } (i,j)^{\text{th}} \text{ element of } \widehat{\boldsymbol{\Omega}}. \end{aligned}$$

Given these definitions, note that $\hat{\sigma}_k^2 \equiv \hat{\omega}_{kk}$. Then let $A_{n,k} = -\hat{\sigma}_k\sqrt{2\log\log n}$, so that $\hat{\boldsymbol{\mu}}$ is defined as a vector with the k^{th} element $\hat{\mu}_k = \bar{d}_k \mathbb{1}_{(\sqrt{n}\bar{d}_k \leq A_{n,k})}$, where $\mathbb{1}_{(X)}$ denotes the indicator function, contingent on event X.

With all the necessary terms defined, the process for the step-SPA method is now described. The step-SPA method tests the following hypotheses:

$$H_0^k : \mu_k \leq 0, \quad k = 1, \ldots, m$$

The test statistics used for the step-SPA method are $\bar{\boldsymbol{d}}_1, \ldots, \bar{\boldsymbol{d}}_m$. Bootstrapping is used to find the null distribution against which these statistics can be checked. The exact bootstrapping method used by both P.-H. Hsu, Y.-C. Hsu and Kuan (2010) and this dissertation is the stationary bootstrap procedure of Politis and Romano (1994). This method generates B resamples of $d_{k,1}, \ldots, d_{k,n}$ for each $k = 1, \ldots, m$. For a single resample of the k^{th} set of d values, let $d^*_{k,i}$ denote the i^{th} value of the resample, for $i = 1, \ldots, n$. The bootstrap procedure chooses $d^*_{k,i} = d_{k,\eta_i}$, where the indices η_i are chosen according to the following rules. For $i = 1$, the value of η_i is chosen randomly, with equal probability, from $\{1, \ldots, n\}$. For $i = 2, \ldots, n$, the value of η_i can take on one of two values, according to a parameter $Q \in [0, 1)$. With probability Q, $\eta_i = \eta_{i-1} + 1$. With probability $1 - Q$, η_i is chosen randomly, with equal probability, from $\{1, \ldots, n\}$. For each resample the value $\bar{d}^*_k = \frac{\sum_{i=1}^n d^*_{k,i}}{n}$ can be calculated, and the vector $\bar{\boldsymbol{d}}^* = (\bar{d}^*_1, \ldots, \bar{d}^*_m)'$ can then be defined. The bootstrap procedure, therefore, results in an empirical distribution of $\bar{\boldsymbol{d}}^*$ with B realisations.

After the bootstrapping procedure, the critical value for a single step in the step wise algorithm is then found using the pre-specified level α_0, and the bootstrapped probability measure P^*. This critical value is defined as $\hat{q}^*_{\alpha_0} = \max(\hat{q}_{\alpha_0}, 0)$, with $\hat{q}_{\alpha_0} = \inf\{q \mid P^*[\sqrt{n} \max_{k=1,\ldots,m}(\bar{d}^*_k - \bar{d}_k + \hat{\mu}_k) \leq q] \geq 1 - \alpha_0\}$. The full algorithm for the step-SPA test then follows the following steps:

1. Sort $\bar{d}_k$ in descending order. Let $\bar{d}_{(1)} \geq \ldots \geq \bar{d}_{(m)}$ denote these sorted values.

2. Find s as the smallest value for which $\bar{d}_{(s)} > \hat{q}^*_{\alpha_0}$ is true. If there is no such value for s, then none of the hypotheses can be rejected, and the algorithm stops.

3. If a value for s can be found, reject the hypotheses which relate to $\bar{d}_{(1)}, \ldots, \bar{d}_{(s)}$, and remove $\bar{d}_{(1)}, \ldots, \bar{d}_{(s)}$ from the data.

4. Repeat steps 1 - 3, using the reduced data set and a new $\hat{q}^*_{\alpha_0}$ critical value based on a bootstrap of said reduced data set.

For this dissertation, the value of B is set to 500, and the value of Q is set to 0.9. This is in line with the values used by P.-H. Hsu, Y.-C. Hsu and Kuan (2010).

A final point to discuss with regard to the implementation of the step-SPA method in this dissertation is the use of studentised test statistics. While all three of the papers which improve on the WRC method (Hansen, 2005; Romano and Wolf, 2005; P.-H. Hsu, Y.-C. Hsu and Kuan, 2010) suggest that studentisation can be used to improve the power of their methods, the results of their papers do not show this to be a major factor. P.-H. Hsu, Y.-C. Hsu and Kuan (2010) find that studentisation makes a very small improvement to the power of their test, and Romano and Wolf (2005) find that studentisation can both increase or decrease the power of their test, depending on other factors in the test. Given these results, this dissertation does not use studentised test statistics, as they would require an additional layer of analysis in which the non-studentised step-SPA method is compared to the studentised version. This layer of analysis would detract from the primary objective of this dissertation - comparing the step-SPA method to the step-MCP method - and was, therefore, judged to be outside the scope of the research.

3.1.4 Step-MCP Method

This section first covers the implementation of the single step MCP method. The adjustment to include a step wise algorithm is described thereafter.

The MCP method is based on the idea that a profitable trading strategy needs to make informed decisions, with the opposite of informed decisions being random decisions. This leads to the null hypothesis for the MCP method which is that the best performing strategy from all the strategies tested is no better than the best performing strategy would be if all the strategies considered were based on random trading decisions. In order to describe the approach for the MCP method, the following definitions are required:

$$
\begin{aligned}
m &: \text{ Number of strategies being tested.} \\
k &: \text{ Subscript for the strategy number, } k = 1, 2, \ldots, m. \\
n &: \text{ Number of periods for each strategy.} \\
t &: \text{ Subscript for the period number, } t = 1, 2, \ldots, n. \\
r_t &: \text{ The asset return in period } t. \\
s_{k,t} &: \text{ The strategy signal for strategy } k \text{ in period } t.
\end{aligned}
$$

where a strategy signal can be one of three values, 1, 0 or -1, corresponding to a buy, neutral or sell signal respectively. This leads to fact that the strategy return for strategy k in period t can be defined as $d_{k,t} = r_t \times s_{k,t}$. This in turn allows for the mean return for strategy k to be defined as $\bar{d_k} = \frac{\sum_{t=1}^{n} d_{k,t}}{n}$.

In order to test the hypothesis for the MCP method, a null distribution is found using Monte Carlo permutations. The process for a single permutation p is as follows. First, let $\boldsymbol{r} = (r_1, \ldots, r_n)'$ be the vector of original asset

returns, and $\boldsymbol{r}^*$ be the vector of asset returns used for this permutation. To populate $\boldsymbol{r}^*$, values are drawn randomly, with equal probability, from $\boldsymbol{r}$ without replacement. This means that $\boldsymbol{r}^*$ now has n elements, with the t^{th} element defined as r_t^*. Second, the strategy returns for this permutation are then calculated as $d_{k,t}^* = r_t^* \times s_{k,t}$. Third, the mean return for each strategy for this permutation is then calculated as $\bar{d}_k^* = \frac{\sum_{t=1}^{n} d_{k,t}^*}{n}$. Finally, the result for the single permutation p is calculated as $l_p = \max_{k=1,\ldots,m}(\bar{d}_k^*)$. This process is repeated for P permutations. The empirical null distribution is found by sorting the values $l_1, \ldots, l_P$ in decreasing order. Let $l_{(1)} \geq l_{(2)} \geq \ldots \geq l_{(P)}$ denote these sorted values. The critical value based on a level of α_0 is then found, using the null distribution, as $q_{\alpha_0} = l_{(\alpha_0 \times P)}$. The null hypothesis for the MCP method is then rejected if $\max_{k=1,\ldots,m}(\bar{d}_k) > q_{\alpha_0}$.

As was mentioned in section 2.4.2, the MCP method can only identify a single strategy as being profitable. Therefore, this dissertation extends the method to include a step wise algorithm, similar to how Romano and Wolf (2005) extended the WRC method. The main difference between the algorithm of Romano and Wolf (2005) and the algorithm used for the step-MCP method is the fact that the former can reject multiple hypotheses in a single step, while the latter only rejects at most a single hypothesis per step. This difference is necessary due to the different structure of the null hypotheses used by stepwise methods based on the WRC and the MCP method. The stepwise WRC-based methods test the joint hypotheses that each strategy's mean is less than zero, therefore, allowing a decision to be reached for each strategy in every step of the algorithm. On the other hand, the MCP method's null hypothesis only concerns the best strategy, and as a result a decision can only be made regarding the best strategy at each step of the algorithm.

The step wise algorithm for the step-MCP iterates through the following two steps:

1. Use the MCP method to determine whether or not the performance of the best strategy is statistically significant. If it is not, the procedure stops.

2. If the performance of the best strategy is found to be significant, remove the strategy from the set of strategies being considered and repeat step one using the revised set.

The step wise algorithm works with the MCP method due to the fact that the null hypothesis tests against the assumption that *all* strategies under consideration are based on random decisions. This can be illustrated using an example, where strategy A is the best strategy under consideration, and strategy B is the second best. For ease of discussion, this example uses the shorthand of a strategy being rejected to refer to the hypothesis related to a specific strategy being rejected. In the example, if strategy A is not rejected, then there is no need to carry out a further test for strategy B, as it would involve testing a worse test statistic against the same null distribution, a redundant exercise. However, if strategy A is rejected, then it must be excluded from further tests, as it is known not to be based on random decisions, and would therefore distort the null distribution. The strategy B then becomes the best strategy in the revised set of strategies under consideration, and the MCP method can be carried out again. This will result in a different null distribution compared to the first step, allowing for the possibility of rejecting strategy B.

For this dissertation, the number of permutations, P, is set to two thousand. This value is chosen as it is large enough to produce a valid test, but

small enough to keep computational times manageable.

3.2 Simulation

The analysis for this paper will be carried out using a Monte Carlo simulation. As has already been discussed, this entails carrying out a large number of simulated investigations, in order to observe the performance of the methods being analysed. This section details the methodology used to implement this simulation. Due to the fact that both WRC and MCP-based methods are being compared, the basic set up of the simulation is the same as that of Perumal and Flint (2018). However, in order to address the areas of their methodology which are not appropriate in the context of TA investigations, as noted in section 2.5, some adjustments are made to the details of their approach. The approach requires two types of data series to be generated, as well as a number of parameters to be defined. The first type of data series is a single set of returns to be used as the underlying asset's returns. The second type of data series is multiple sets of trading signals , one for each strategy in the simulation. The methods used to generate these data series are described in sections 3.2.1 and 3.2.2 below. The parameters used are defined, and their chosen values discussed in section 3.2.3 below.

The simulations were run using the R programming language (R Core Team, 2019), with an object orientated approach facilitated by the *R6* package (Chang, 2019). Monte Carlo simulation is extremely computationally intensive in general, and due to the fact that this dissertation required multiple simulations to be run, extensive use was made of parallel computing packages for R (Microsoft and Weston, 2017; Microsoft Corporation and Weston, 2018). In particular, the *doRNG* package of Gaujoux (2018) was used to ensure that the randomisation for the simulations was both independent and

replicable across the parallel runs. Implementing the simulation study in this manner also required significant computational resources. These were kindly provided by the University of Cape Town's (UCT) ICTS High Performance Computing (HPC) team.

3.2.1 Generating The Return Series

When generating a series of returns, the sophistication of the approach chosen will depend on the trade-off between pragmatic implementation and accuracy in representing actual returns. For example, a simple, pragmatic approach for generating returns would be to draw a sample of normally distributed values. However, it is well known that returns are autocorrelated, therefore, the generated series would not be a good representation of reality. This problem is made more complex by the fact that the true nature of returns is not known, therefore, perfect accuracy when simulating returns is not possible. As a result, there are many models and approaches which can be used to generate returns for a simulation, and there is no consensus on which method is preferred.

In light of this uncertainty, it was decided that this dissertation should base its approach on the method used by P.-H. Hsu, Y.-C. Hsu and Kuan (2010). They generate the t^{th} return as:

$$r_t = c + \gamma r_{t-1} + \epsilon_t \tag{1}$$

where c and γ are constants, and ϵ_t is drawn from a $N(0, \sigma^2)$ distribution independently for each t. This method allows for the autocorrelation and overall mean of the return series to be controlled, and assumes that the variance of returns is constant over time.

It is well known, however, that the variance of equity returns is not constant and that it also displays clustering (Nelson, 1991). Therefore, a second

method for generating returns was also used, in order to test the methods under a variance assumption which is more in line with what might be expected from actual asset returns. This method is the same as that of P.-H. Hsu, Y.-C. Hsu and Kuan (2010), however, instead of using constant variance, a generalized autoregressive conditional heteroskedasticity (GARCH) model (Bollerslev, 1986) is used to set the variance for each period. Using a GARCH model changes the value for ϵ_t used in equation 1. A GARCH(p, q) model sets ϵ_t to be distributed according to $N(0, \sigma_t^2)$ where σ_t^2 is calculated as follows:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

where $p \geq 0$, $q > 0$, $\omega > 0$, $\alpha_i > 0$, $\beta_j > 0$, $i = 1, \ldots, q$ and $j = 1, \ldots, p$. The parameter values used for the GARCH model in this dissertation are discussed in section 3.2.3. When labelling the different simulations in the results section, the simulations which use the default method are labelled as *hsu*, while the simulations which use the altered method are labelled as *garch*.

3.2.2 Generating The Trading Signal Series

The set of trading signals need to be generated in a way which allows for the number of profitable, unprofitable and neutral strategies to be controlled. The profitability of a strategy can be controlled by adjusting the number of correct signals generated for said strategy. A correct signal is a signal which either matches a buy instruction with a positive return, or a sell signal with a negative return. Perumal and Flint (2018) achieved this by setting the number of correct decisions which a rule makes as a proportion of the total number of decisions made. Therefore, if a rule is profitable, the proportion is set to greater than fifty per cent, if it is unprofitable, the proportion is set to

less than fifty per cent, and if the rule is neutral, the proportion is set to fifty per cent exactly. This approach includes an implicit assumption that if the rule does not give the correct signal, then it gives an incorrect signal and vice versa. As was discussed in section 2.5, this assumption is not appropriate in the context of TA investigations.

To account for this, this dissertation will generate trading signals using a different approach based on a strength parameter. This strength parameter will control how likely it is that a signal is set to the correct value at each time point. For profitable strategies the correct value is a buy signal if the return for the period is positive and a sell signal if the return for the period is negative. For unprofitable strategies the correct value is a sell signal if the return for the periods is positive and a sell signal if the return for the period is negative. For neutral strategies, the strength parameter is simply set to zero, which has the effect of never setting a signal to the correct value. While the strength parameter approach is similar to that of Perumal and Flint (2018), it allows for the following important difference. When a signal is not set to be correct, instead of automatically setting the signal to the opposite value, it is set to any one of the three signal options with equal probability. This approach allows for the profitability of the generated rules to be controlled, without requiring the assumption that if a signal is not chosen to be correct, then it must be incorrect and vice versa.

The strength based approach for a profitable strategy can be formulated as follows. Let the strength parameter be s, with possible values $0 \leq s \leq 1$. Let the signal for a given time t be g_t and the period return for this time be r_t. For the signal value, 1 is equivalent to a buy signal, -1 is equivalent to a sell signal, and 0 is equivalent to a neutral signal. With probability s, g_t is

set according to the formula

$$g_t = \frac{r_t}{|r_t|} \tag{2}$$

where $|.|$ denotes the absolute value. With probability $(1-s)$, g_t has an equal chance of being set to 1, 0 or -1.

The above description details how profitable strategies are generated, however, this approach can easily be adapted to produce unprofitable or neutral strategies. For unprofitable strategies, equation 2 is replaced with

$$g_t = -\frac{r_t}{|r_t|} \tag{3}$$

For neutral strategies the strength parameter is simply set to zero.

3.2.3 Analysis Parameters

As has been mentioned, Monte Carlo simulations are extremely computationally intensive. As a result, some choices regarding the simulation parameters were influenced by the need to reduce the computational burden of the simulations. These choices were made in areas which would be least likely to influence the results of the simulation. However, the fact that these choices were made, means that the results were still influenced. This provides scope for future research to address these choices, and the limits which they placed on the investigation.

The parameters for the simulation can be split into two groups. The constant parameters, which remain the same across all the simulations, and the variable parameters, which are altered for different simulations to test the robustness of the various methods. In order to keep the number of simulations to be run at a manageable number, not all combinations of the variable parameters are considered. Instead, a base-case set of values was chosen for the variable parameters. Simulations are run on this base-case, as

well as the combinations arising from substituting in the alternate values of the variable parameters one at a time. In addition to limiting the number of combinations being tested, this also allows for the effects of the alternate values to be tested in isolation, as all other parameters are held constant.

The constant parameters are now introduced, with a summary of their values presented in table 2. The first constant parameter is the number of replications per simulation. This parameter has the largest impact on the runtime of the simulation, but due to the fact that it needs to be sufficiently large for useful inference to be drawn from the simulations, it cannot be set too low. Taking this into account, a value of 520 was chosen. This value is in line with the value of 500 used by P.-H. Hsu, Y.-C. Hsu and Kuan (2010) in their simulations, adjusted upward to be a multiple of 40, the number of CPUs on a single node of the UCT HPC clusters. The second constant parameter is the number of periods which each simulation will consider. This is set to 1 250, to represent approximately five years of daily data. The third constant parameter is the constant c for the return generating process. This value is set to zero, so that the returns generated for a neutral strategy will have a mean of zero. The fourth constant parameter is the σ value for the hsu return generation method. This value is set to 0.005, the same value as that used by P.-H. Hsu, Y.-C. Hsu and Kuan (2010) for their simulations. The remaining constant parameters are for the GARCH model used to generate returns. Their values are chosen as follows: First, the order of the GARCH model is set to $(1, 1)$. This choice is made due to the fact that the GARCH$(1, 1)$ model is the most popular version of the GARCH model used for financial time series (Poon and Granger, 2003). Second, the α value is set to 0.05, and the β value is set to 0.9. These values are chosen to be within the range of expected parameter values for a GARCH model on

daily data, as suggested by Alexander (2008). Finally, the ω value is set to 0.00000125. This value is chosen so that the unconditional variance of the GARCH model is 0.000025, the squared value of the σ parameter for the hsu return generation method.

Table 2: Summary of constant parameters.

Parameter	Value
Replications	520
Periods	1250
c	0
σ	0.005
GARCH(p, q)	(1,1)
GARCH: α	0.05
GARCH: β	0.9
GARCH: ω	0.00000125

The variable parameters are now introduced, with a summary of their values in table 3. Each one is assigned a label for reference during the results section of this dissertation. Furthermore, the parameter values marked with an asterisk are those used for the base-case simulations. The first variable parameter is the number of strategies considered, labelled M. This parameter is varied due to the fact that the severity of the DM bias increases as the number of strategies increases (Aronson, 2011; Perumal and Flint, 2018). Therefore, it is useful to see how the different methods perform at varying levels of strategies considered. The three values considered for this parameter are 50, 100 and 250*. While the number of strategies tested in practice will often be much larger than this, this parameter has a significant effect on

the computation time of the simulations. Therefore, a compromise has been made between values which will demonstrate the capabilities of the methods tested and values which allow for reasonable computation time.

The second variable parameter is the constant *gamma* for the return generating process. This parameter controls the autocorrelation in the generated returns and is labelled *Correlation*. The correlation parameter is varied due to the fact that MCP-based methods can become anti-conservative in the presence of autocorrelation, as was discussed in section 2.4.2. Therefore, it is necessary to analyse how robust the various methods are to changes in this parameter. The three values considered for this parameter are 0, 0.01* and 0.1. The choice of zero is self-explanatory, the choice of 0.01 is in-line with the value used by P.-H. Hsu, Y.-C. Hsu and Kuan (2010) and the choice of 0.1 is in-line with the value suggested by Masters (2006) as an extreme value for this parameter.

The third variable parameter is the number of profitable strategies in the overall group being considered, labelled $M1$. This parameter is varied in order to analyse whether or not different values affect the power of the different methods. The three values considered for this parameter are 1, 10* and 50. These values are chosen to sufficiently test the robustness of the different methods, while still being within the range of values which could occur in practice.

The fourth variable parameter is the number of unprofitable strategies in the overall group being considered, labelled $M3$. The three values considered for this parameter are 0, 10*, 50. The motivations for both varying this parameter and for the values chosen are the same as for the $M1$ parameter.

The fifth variable parameter is the strength for the trading signals generated, labelled *Strength*. This parameter is varied in order to analyse how sens-

itive the different methods are to the strength of the non-neutral strategies considered. The four values considered for this parameter are 0.05, 0.1, 0.15* and 0.25. The motivation for the choice of these values is also the same as for the $M1$ parameter.

The sixth variable parameter is the significance level used for the methods, labelled *alpha*. This parameter is varied in order to analyse whether or not the different methods still control for the DM bias correctly at different levels of significance. The two values considered for this parameter are 0.01 and 0.05*. These values are chosen to be in line with the significance levels chosen for most research in the financial field.

The final variable parameter is the method used to generate returns, labelled *Return.Generation*, it has already been discussed in section 3.2.1 and takes on the values *hsu* or *garch**.

Table 3: Summary of variable parameters.

Parameter	Base Value	Alternate Values
M	250	50, 100
γ	0.01	0, 0.1
$M1$	10	1, 50
$M3$	10	0, 50
Strength	0.15	0.05, 0.1, 0.25
α	0.05	0.01
Return.Generation	*garch*	*hsu*

3.3 Analysis of Performance

The performance of the various methods will be analysed using the three metrics defined in the literature review, namely the FWE, FD and AR rates. These rates give an idea of the probability that a rule will be chosen when it shouldn't as well as the probability that a profitable rule will not be identified. Therefore, this allows for both the power and the accuracy of the models to be tested. A good model will be one which has high power, while still accurately controlling the FWE rate. If only the FD rate is controlled, this is less ideal, but still an advantage over not controlling for the DM bias at all.

Both the FWE and AR rates are already designed to provide information from a simulation study. The FD rate, on the other had, is calculated for each replication of a simulation. Therefore, in order to get an idea of how well the FD rate was controlled across multiple replications, a new measure is defined, the Average False Discovery (AFD) rate. In order to calculate the AFD rate, the FD rate for each replication is calculated, and the average of these individual FD rates is then found to give the AFD value for the full simulation.

4 Results

This section presents the results of the simulation study. The tables presented were compiled with the aid of the *x-table* package for R (Dahl *et al.*, 2019), while the figures presented were generated using the *ggplot2* and *gridExtra* packages for R (Wickham, 2016; Auguie, 2017). For the sake of brevity, the step-MCP and step-SPA methods are referred to simply as the MCP and SPA methods respectively.

4.1 Base-case Simulations

The results of the base-case simulations for all the methods are displayed in table 4. The power for all four methods is very good, with all methods having an AR rate of more than 0.9. The results for the measures of type-I error are less straightforward. Both the T and the BH methods have a FWE rate much higher than 0.05, indicating that the methods fail to sufficiently control for the DM bias when considering the FWE rate. On the other hand, the MCP and SPA methods both have FWE rates below 0.05, indicating that these methods do sufficiently control for the DM bias in this regard. Furthermore, the AFD rates for the MCP and SPA methods are less than 0.01, while the rate for the BH method is less than 0.05 and the rate for the T method is much greater than 0.05, indicating that only the T method does not control for the DM bias when considering the FD rate.

Table 4: Performance measures for the base-case simulations.

	AR	FWE	AFD
MCP	0.9425	0.0442	0.0042
SPA	0.9344	0.0346	0.0035
BH	0.9837	0.3865	0.0445
T	0.9998	1.0000	0.5226

These results are as expected. The T method performs extremely well at identifying profitable strategies, with an AR rate of 0.99. However, due to the fact that it does not account for the DM bias, it is also extremely anti-conservative, with a FWE rate of one and AFD rate of approximately 0.5. The BH method also performs extremely well at identifying profitable strategies, with an AR rate of 0.98. It also controls the FD rate as per its

design, with an AFD rate of 0.044. However, the fact that the method only seeks to control the FD rate of the test, means that it is anti-conservative when looking at the FWE rate, with a value of approximately 0.38. The two randomisation based methods, the MCP and SPA methods, have lower power than the above two methods, with AR rates of 0.94 and 0.93 respectively. However, these levels of power are still acceptable, and they are justified by the fact that both methods control the FWE rate, as per their design, with values of 0.04 and 0.03 respectively. The AFD rates for these methods are both substantially smaller than 0.05, which is to be expected, as control of the FWE rate is a stricter constraint than control of the FD rate.

The one noteworthy result from the base case is the fact that the improved MCP method performs well, demonstrating the ability to identify multiple profitable strategies while controlling the FWE rate. Furthermore, its performance in terms of power is in line with that of the SPA method.

4.2 Alternate-case Simulations

In order to keep the discussion of the alternate-case simulations focused on relevant results, some methods are excluded from the discussion of certain performance measures. The T method is excluded from the discussion of all three performance measures, as its results simply show evidence of the DM bias through all iterations. The BH Method is excluded from the discussion of the FWE rate, as the method does not seek to control the measure and the results simply show this to be true. Finally, when discussing the AFD rate, only the BH method is discussed. This is due to the fact that the AFD rates for the SPA and MCP methods are always small enough to not provide valuable insight. The full results for the alternate-case simulations of each method are presented in the appendices.

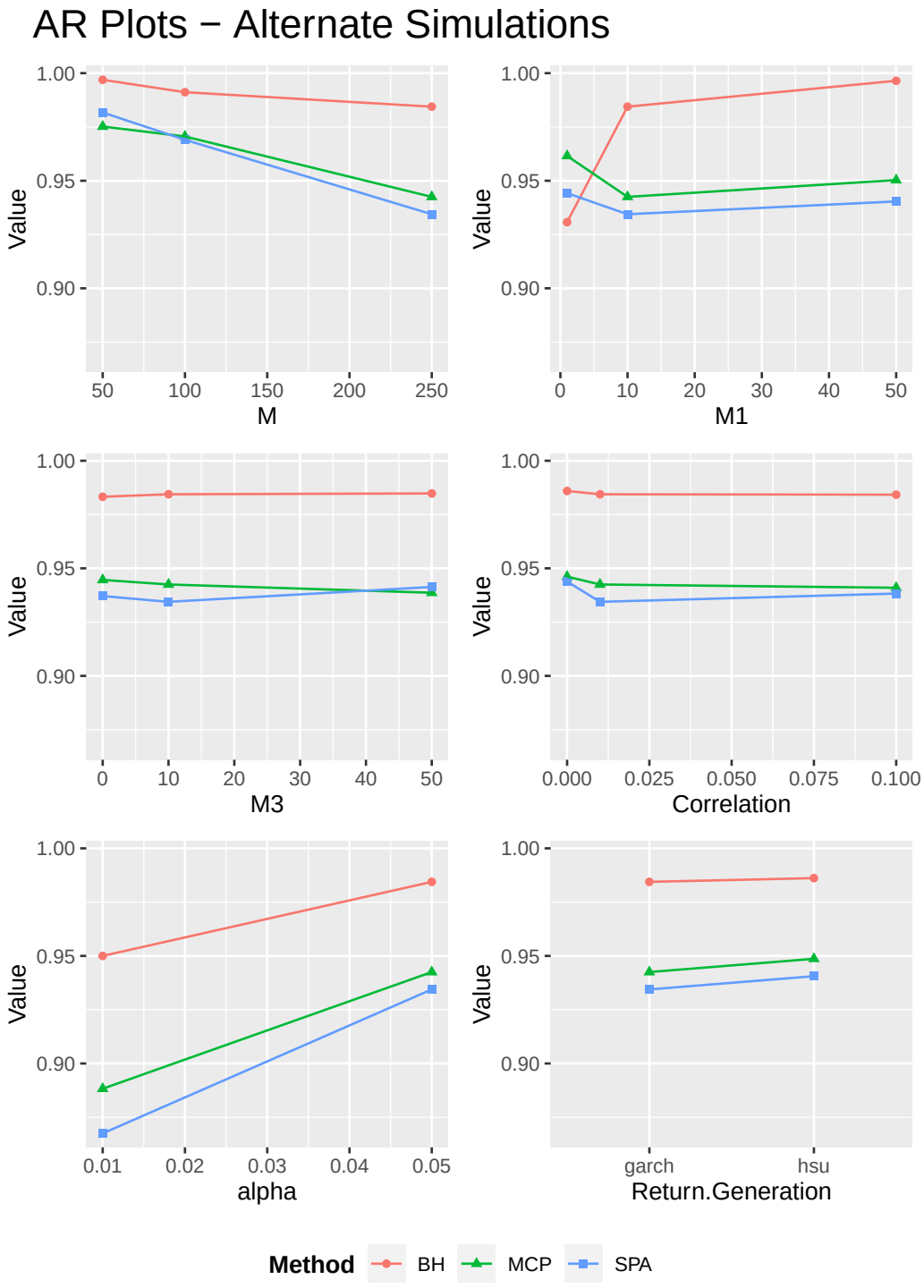

Figure 1: Plots of the AR values for all alternate simulations, except those of the strength parameter.

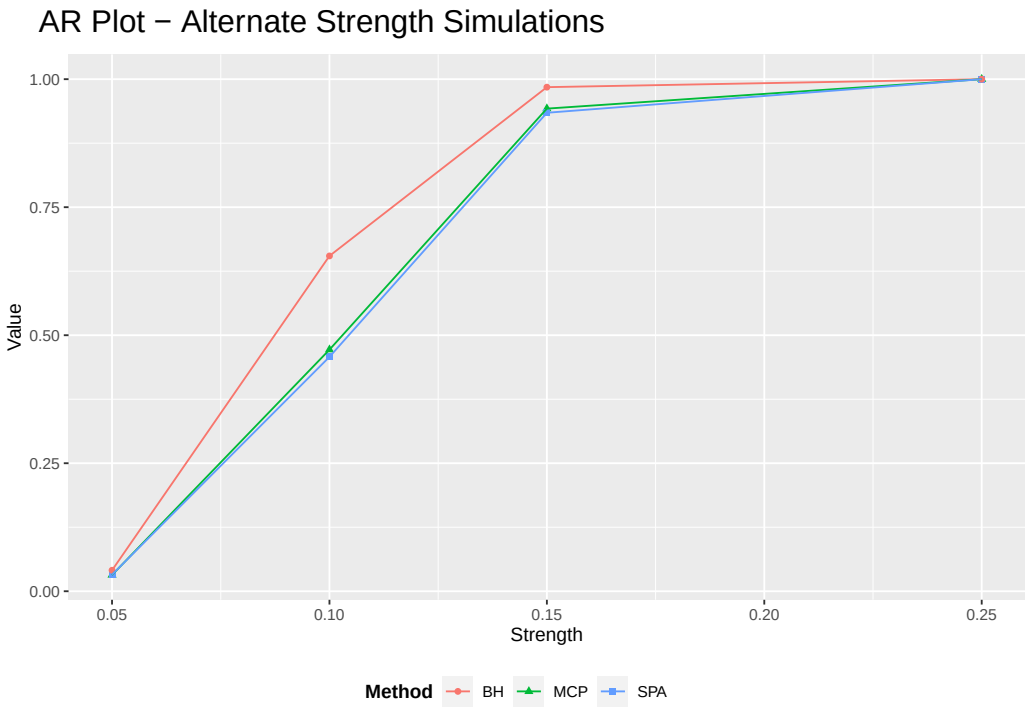

Figure 2: Plot of the AR values for the alternate strength simulations.

The first performance measure considered in this discussion is the AR rate, which provides an indication of a method's power. The AR values for all of the alternate simulations except for those on the strength parameter are displayed in figure 1. The values for the alternate strength simulations are displayed separately, in figure 2, due to the vastly different y-axis scale for the figure. When considering the three methods together, it can be seen from the figures that the level of power is relatively stable for changes in $M1$ (except the BH method, as discussed below), $M3$, the level of correlation and the method for generating returns. In contrast to this, the level of power decreases slightly as M increases, and as alpha decreases. The level of power also falls close to zero as strength decreases. The decrease in power for these three parameters is relatively intuitive as a larger group of strategies

being tested, a stricter level for significance and weaker returns for profitable strategies all make it more difficult to identify profitable rules as being significant. The fact that changes in the strength parameter have a much larger effect than changes in any of the other parameters suggests that the level of this parameter is likely to dominate any other concerns when considering the power of an investigation.

When considering the power level of the individual methods, the first point to note is that the BH method has superior power compared to the other two methods in all but one of the different configurations. This is not entirely unexpected, as the BH method uses a weaker constraint to control for the DM bias, so it should also offer better power. What is noteworthy, however, is the fact that, if the $M1 = 1$ configuration is set aside, the magnitude of the difference in power is relatively large. Furthermore, in the configurations which have a negative impact on power across all three methods, namely weaker strength and a larger M value, the power of the BH method decreases less steeply. These two facts are noteworthy, as they indicate that, before the other performance measures are considered, the BH method is a strong choice for use in TA investigations. As for why the BH method displays increasing power as the $M1$ value increases, while the other methods display relatively stable power, it is likely due to the manner in which the method's design interacts with the simulation methodology. The profitable strategies in the simulation all have the same strength level, which means that they should all have relatively similar p-values. Therefore, by applying a stricter p-value adjustment in cases with fewer profitable rules, the BH method is effectively testing a similar p-value against a smaller adjusted significance level. This explains the observed decrease in power for the lower $M1$ values.

When comparing the two remaining methods to one another, the MCP

method has superior power compared to the SPA method under most of the configurations, however, the difference in power is never of a particularly substantial magnitude. This close relationship between the MCP and SPA methods is particularly noteworthy for the two configurations with a non-zero $M3$ value. This indicates that the MCP method maintains a similar power level compared to the SPA method under these configurations, even though the MCP method isn't specifically designed to maintain power in the presence of poor strategies, as is the case with the SPA method.

The second performance measure considered in this discussion is the FWE rate. Recall that the BH method is omitted from this discussion due to the fact that it does not seek to control the FWE rate, resulting in FWE values which provide no insight other than the fact that they are all significantly high. The FWE values for all of the alternate simulations are displayed in figure 3. If a method is correctly controlling the FWE rate, then its value should be less than or equal to the significance level for the test. On the plots in figure 3 the significance level is shown as a dotted line. From the plots it can be seen that the SPA method successfully controls the FWE rate under all configurations. On the other hand, the MCP method successfully controls the FWE rate for most of the configurations with the exceptions being the $M1 = 1$ configuration, the strength $= 0.05$ configuration, the correlation $= 0.1$ configuration, and the alpha $= 0.01$ configuration. Of these four results, only the high FWE rate for a higher level of correlation is expected. This is due to the fact that, as was discussed in section 2.4.2, the MCP-based methods can become anti-conservative in the presence of autocorrelation.

To understand why the other three configurations also have a high FWE rate, one needs to consider the affect of other base-case parameter values on

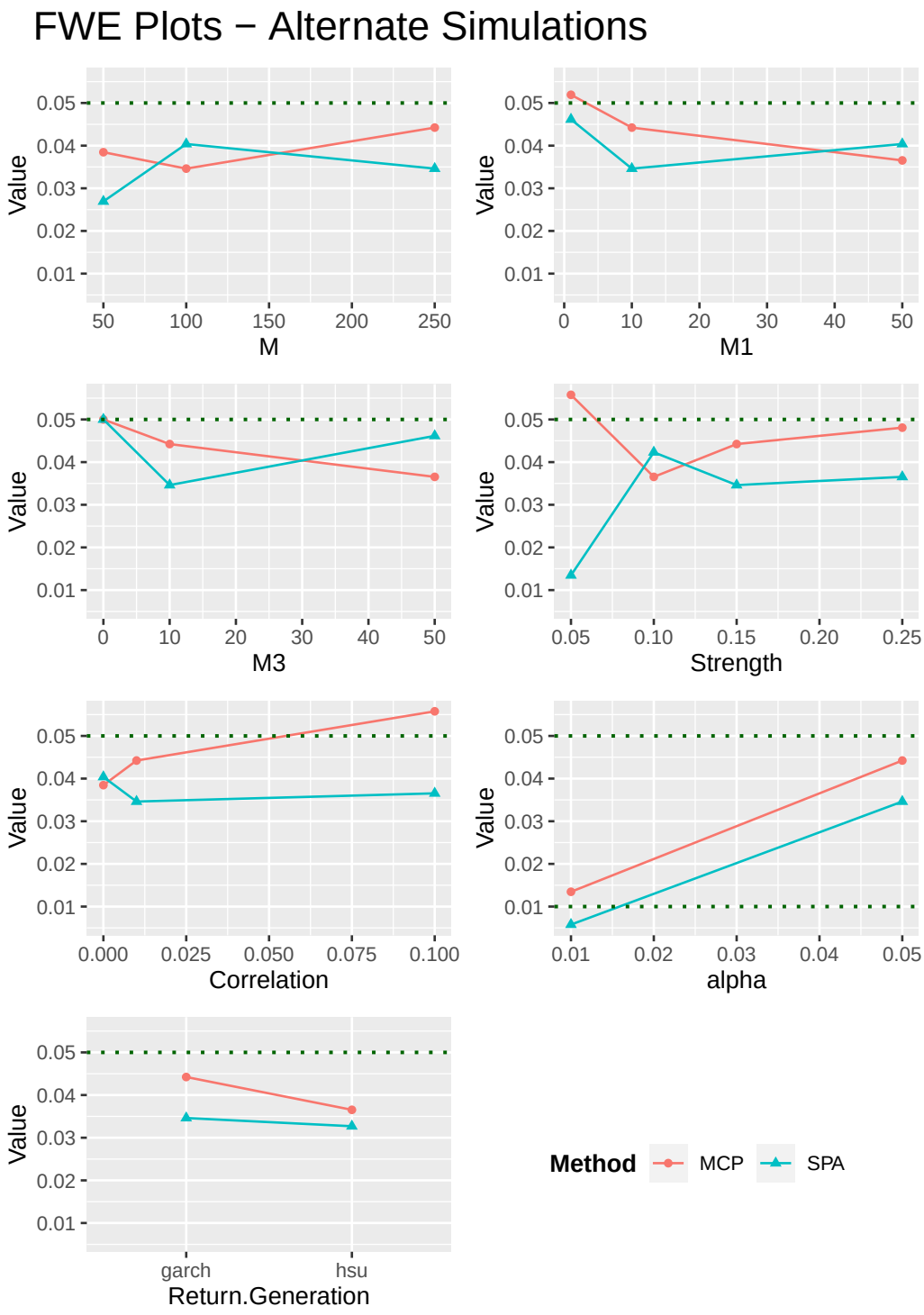

Figure 3: Plots of the FWE values for all alternate simulations.

the FWE rate for the MCP method. Certain parameters, such as the $M3$ value, the $M1$ value and the strength value for the base-case all have the affect of lowering the FWE rate. This means that even though the base-case does have a non-zero correlation value, the increase that this causes in the FWE rate is offset by the effects from the other parameters. However, when the effects from these other parameters are reduced, or a lower significance level is used (resulting in a smaller magnitude for these effects), then the increase caused by the correlation parameter dominates the other effects. This is why the other three configurations display an FWE rate above 0.05. To confirm that this was the cause of the high FWE values, the simulations for the $M1 = 1$ configuration, the strength $= 0.05$ configuration and the alpha $= 0.01$ configuration were rerun, using a correlation value of zero instead of the base-case value of 0.01. The resulting FWE values for the three configurations are 0.0500, 0.0308 and 0.0077, all less than or equal to the relevant significance levels. This provides strong evidence that the only area of concern when considering how well the MCP method controls the FWE rate is the level of autocorrelation.

The third performance measure considered in this discussion is the AFD rate. The AFD values for all of the alternate simulations are displayed in figure 4. If a method is correctly controlling the AFD rate, then its value should be less than or equal to the significance level for the test. On the plots in figure 4 the significance level is shown as a dotted line. From the plots it can be seen that the BH method does correctly control the AFD rate for all configurations except when $M1 = 1$. This is the same configuration under which the BH method displayed an unusually low level of power, relative to its power level for other configurations. Once again, this apparent outlying case for the BH method is likely a result of the simulation methodology. For

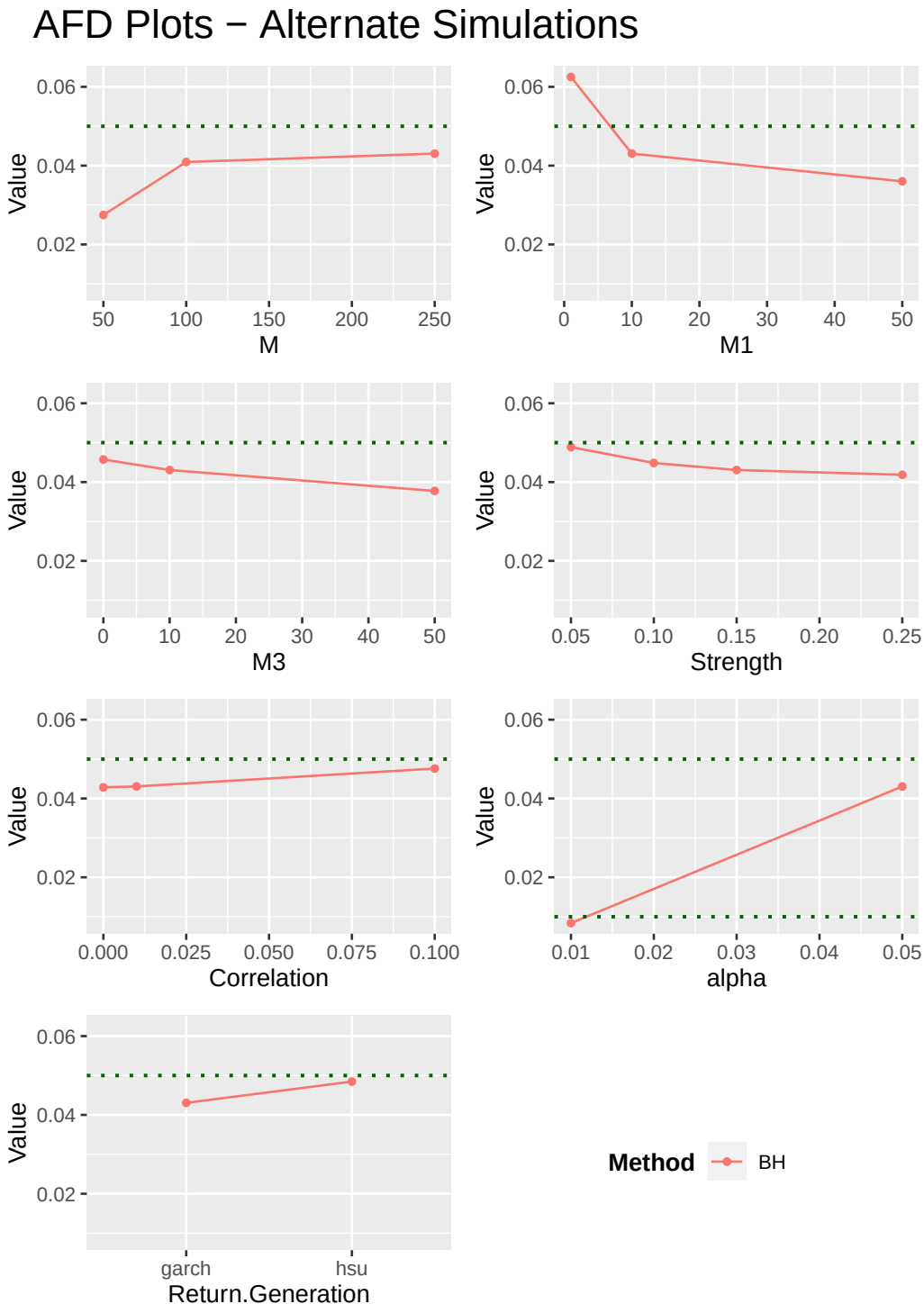

Figure 4: Plots of the AFD values for all alternate simulations - BH method only.

the $M1 = 1$ configuration, the possible values for the FD rate from a single simulation run are only one, half, or zero. Therefore, when taking the average of these rates over only 520 simulations, the AFD value can be inflated above its true value. This effect can also be seen in the AFD values for the other three methods under the $M1 = 1$ configuration. To test whether or not this was the cause of the high AFD value, the $M1 = 1$ configuration of the BH method was rerun, using 1 500 simulation runs. The resulting AFD value is 0.0463, which is further evidence that the simulation methodology is the likely cause for the outlying AFD value observed. Therefore, the results presented in figure 4 can be taken as strong evidence that the BH method successfully controls the AFD rate under all configurations.

4.3 Method Run-time

One aspect of the different methods which previous research has not considered is the relative computational time they require. While this dissertation does not conduct and in-depth investigation in this regard, the run-times of the different methods were recorded for the various simulations. A clear disparity between the different methods was observed. This motivated a simple test of the run-times, using a single replication of the base-case simulation for each method. The absolute values recorded during this test are not particularly meaningful, as they will vary according to the resources used to run the simulations. However, insight can be drawn from the times relative to one another. The BH method completed after approximately two seconds. The MCP method completed after approximately three minutes. The SPA method completed after approximately eight minutes. The extreme difference between the BH method and the other two methods is to be expected, considering the fact that the BH method is not a randomisation method. The

fact that the MCP method has a run-time less than half that of the SPA, is less expected. This difference is significant due to the fact that most TA investigations will typically consider a large number of strategies, requiring a substantial amount of computational time. Being able to complete the investigation using less that half the computational time could be a significant advantage, particularly if computational resources are limited or expensive.

5 Conclusion

TA investigations are particularly susceptible to the effects of the DM bias. This dissertation compared three methods which can be used to conduct TA investigations while controlling for the effects of the DM bias. It was found that all three methods controlled for the DM bias as their designs intended, with the possible exception of the step-MCP method, which showed evidence of becoming anti-conservative when testing strategies based on returns with a high level of autocorrelation. The power for all three methods displayed some common trends. When the number of rules considered increased, the power of all three methods decreased. This suggests that the common practice in TA investigations of testing as many strategies as possible may be less productive than testing a smaller number of strategies, chosen using a sound logical or theoretical basis. The power for all three methods also decreased as the strength of the profitable strategies decreased. This trend is less worrying for TA investigations due to the fact that failing to identify weakly profitable rules is not detrimental as they will generally not remain profitable once considerations such as trading costs and liquidity constraints are taken into account.

When considering the methods on an individual basis, the following con-

clusions can be drawn. The BH method successfully controlled the FD rate. While this is a less strict constraint compared to the FWE rate, it still allows for useful results to be obtained from the investigation. If the method identifies zero or a small number of profitable strategies, then it is likely that further investigation is not warranted. However, if the method identifies a relatively large number of profitable strategies, then there is a very good chance that $1 - \alpha$ per cent of these strategies are true profitable strategies. If a trader is happy to use all of the identified strategies and absorb the losses from the small incorrectly identified per cent, then they can use the results as is. If a trader wishes to identify the profitable strategies more accurately, they can then use one of the more complicated and time consuming randomisation methods, so long as they use the method on all the strategies originally considered and not only the strategies identified as being profitable by the BH method. This approach is particularly appealing given the fact that the run time for the BH method is a minute fraction of that of the other two methods.

The step-MCP and step-SPA methods perform very similarly across most of the different configurations tested. This means that the step-MCP method also addresses the two main drawbacks of the WRC method, which the step-SPA method was specifically designed to address. Therefore, it can be seen that the extension of the MCP method to the step-MCP method successfully achieved the goal of identifying multiple profitable strategies. Furthermore, the step-MCP method does this while relying on a weaker set of assumptions and with a significantly faster run time. These facts provide strong evidence for the potential of the step-MCP method, however, the fact that it becomes anti-conservative when the asset returns are autocorrelated is a concern. This means that in its current state the method can only be reliably used to test

TA strategies based on assets which it can be safely assumed will not have autocorrelated returns. The further extension of the step-MCP method to address this drawback is a potential area for further research.

Given the concern with the step-MCP method and the fact that the BH method does not control the FWE rate, the step-SPA method is, therefore, the only method which fully controls for the DM bias under all configurations. While it does rely on a stricter set of assumptions and require more computational time to run, it is still the best option for a robust method which accurately accounts for the effects of the DM bias.

References

Alexander, C. (2008). *Market risk analysis. pricing, hedging and trading financial instruments*. Chichester, England: John Wiley & Sons.

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

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. URL: https://CRAN.R-project.org/package=gridExtra.

Benjamini, Y. and Y. Hochberg (1995). 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bollerslev, T. (Apr. 1986). 'Generalized autoregressive conditional heteroskedasticity'. In: *Journal of Econometrics* 31.3, pp. 307–327. DOI: 10.1016/0304-4076(86)90063-1.

Chang, W. (2019). *R6: Encapsulated Classes with Reference Semantics*. R package version 2.4.0. URL: https://CRAN.R-project.org/package=R6.

Chordia, T., A. Goyal and A. Saretto (Aug. 2017). *p-Hacking: Evidence from Two Million Trading Strategies*. en. SSRN Scholarly Paper 17-37. Swiss Finance Institute. DOI: 10.2139/ssrn.3017677.

Covel, M. (2004). *Trend Following: How Great Traders Make Millions in Up or Down Markets*. First. FT Press. ISBN: 0131446037.

Dahl, D. B. *et al.* (2019). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4. URL: https://CRAN.R-project.org/package=xtable.

Fernández-Rodriguez, F., C. González-Martel and S. Sosvilla-Rivero (Oct. 2000). 'On the profitability of technical trading rules based on artificial

neural networks:: Evidence from the Madrid stock market'. In: *Economics Letters* 69.1, pp. 89–94. ISSN: 0165-1765. DOI: 10.1016/S0165-1765(00)00270-6.

Gaujoux, R. (2018). *doRNG: Generic Reproducible Parallel Backend for 'foreach' Loops*. R package version 1.7.1. URL: https://CRAN.R-project.org/package=doRNG.

Hansen, P. R. (Oct. 2005). 'A Test for Superior Predictive Ability'. en. In: *Journal of Business & Economic Statistics* 23.4, pp. 365–380. ISSN: 0735-0015, 1537-2707. DOI: 10.1198/073500105000000063.

Harvey, C. R. and Y. Liu (Aug. 2014). 'Evaluating Trading Strategies'. en. In: *The Journal of Portfolio Management* 40.5, pp. 108–118. DOI: 10.2139/ssrn.2474755.

Hsu, P.-H., Y.-C. Hsu and C.-M. Kuan (June 2010). 'Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias'. In: *Journal of Empirical Finance* 17.3, pp. 471–484. ISSN: 0927-5398. DOI: 10.1016/j.jempfin.2010.01.001.

Hsu, P.-H. and C.-M. Kuan (Oct. 2005). 'Reexamining the Profitability of Technical Analysis with Data Snooping Checks'. en. In: *Journal of Financial Econometrics* 3.4, pp. 606–628. ISSN: 1479-8409. DOI: 10.1093/jjfinec/nbi026.

Lakonishok, J., A. Shleifer and R. W. Vishny (1994). 'Contrarian Investment, Extrapolation, and Risk'. en. In: *The Journal of Finance* 49.5, pp. 1541–1578. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.1994.tb04772.x.

Lo, A. W., H. Mamaysky and J. Wang (2000). 'Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation'. en. In: *The Journal of Finance* 55.4, pp. 1705–1765. ISSN: 1540-6261. DOI: 10.1111/0022-1082.00265.

Masters, T. (2006). 'Monte-Carlo Evaluation of Trading Systems'. URL: https://www.evidencebasedta.com/montedoc12.15.06.pdf (visited on 11/06/2019).

Menkhoff, L. (Nov. 2010). 'The use of technical analysis by fund managers: International evidence'. In: *Journal of Banking & Finance* 34.11, pp. 2573–2586. ISSN: 0378-4266. DOI: 10.1016/j.jbankfin.2010.04.014.

Microsoft Corporation and S. Weston (2018). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.14. URL: https://CRAN.R-project.org/package=doParallel.

Microsoft and S. Weston (2017). *foreach: Provides Foreach Looping Construct for R*. R package version 1.4.4. URL: https://CRAN.R-project.org/package=foreach.

Nelson, D. B. (1991). 'Conditional Heteroskedasticity in Asset Returns: A New Approach'. In: *Econometrica* 59.2, pp. 347–370. DOI: 10.2307/2938260.

Perumal, K. and E. Flint (2018). 'Systematic testing of systematic trading strategies'. In: *Journal of Investment Strategies* 7.3, pp. 29–49. DOI: 10.21314/JOIS.2018.100.

Politis, D. N. and J. P. Romano (1994). 'The Stationary Bootstrap'. en. In: *Journal of the American Statistical Association* 89.428, pp. 1303–1313. DOI: 10.1080/01621459.1994.10476870.

Poon, S.-H. and C. W. J. Granger (June 2003). 'Forecasting Volatility in Financial Markets: A Review'. In: *Journal of Economic Literature* 41.2, pp. 478–539. DOI: 10.1257/002205103765762743.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Romano, J. P. and M. Wolf (2005). 'Stepwise Multiple Testing as Formalized Data Snooping'. en. In: *Econometrica* 73.4, pp. 1237–1282. ISSN: 1468-0262. DOI: 10.1111/j.1468-0262.2005.00615.x.

White, H. (2000). 'A reality check for data snooping'. In: *Econometrica* 68.5, pp. 1097–1126.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Appendices

A Performance Measures for All Methods

Table 5: Performance measures for all T Method simulations

Param	Val	AR	FWE	AFD
Base		0.9998	1.0000	0.5226
M	50	0.9996	0.8058	0.1295
M	100	0.9996	0.9788	0.2693
M1	1	1.0000	1.0000	0.9169
M1	50	0.9997	1.0000	0.1545
M3	0	0.9996	1.0000	0.5290
M3	50	0.9996	1.0000	0.4716
Strength	0.05	0.5288	1.0000	0.6783
Strength	0.1	0.9571	1.0000	0.5343
Strength	0.25	1.0000	1.0000	0.5285
Correlation	0	0.9996	1.0000	0.5242
Correlation	0.1	0.9998	1.0000	0.5271
alpha	0.01	0.9967	0.9058	0.1743
Return.Generation	hsu	0.9996	1.0000	0.5264

Table 6: Performance measures for alternate BH Method simulations

Param	Val	AR	FWE	AFD
Base		0.9844	0.3654	0.0431
M	50	0.9969	0.2500	0.0275
M	100	0.9912	0.3538	0.0409
M1	1	0.9308	0.1173	0.0625
M1	50	0.9965	0.8346	0.0360
M3	0	0.9833	0.3962	0.0457
M3	50	0.9848	0.3423	0.0377
Strength	0.05	0.0410	0.0712	0.0489
Strength	0.1	0.6548	0.2962	0.0448
Strength	0.25	1.0000	0.3635	0.0418
Correlation	0	0.9860	0.3692	0.0428
Correlation	0.1	0.9842	0.4096	0.0476
alpha	0.01	0.9500	0.0827	0.0084
Return.Generation	hsu	0.9862	0.4115	0.0485

Table 7: Performance measures for alternate MCP Method simulations

Param	Val	AR	FWE	AFD
Base		0.9425	0.0442	0.0042
M	50	0.9752	0.0385	0.0038
M	100	0.9706	0.0346	0.0032
M1	1	0.9615	0.0519	0.0276
M1	50	0.9503	0.0365	0.0008
M3	0	0.9446	0.0500	0.0050
M3	50	0.9387	0.0365	0.0038
Strength	0.05	0.0317	0.0558	0.0510
Strength	0.1	0.4717	0.0365	0.0069
Strength	0.25	1.0000	0.0481	0.0044
Correlation	0	0.9462	0.0385	0.0038
Correlation	0.1	0.9410	0.0558	0.0055
alpha	0.01	0.8883	0.0135	0.0013
Return.Generation	hsu	0.9487	0.0365	0.0037

Table 8: Performance measures for alternate SPA Method simulations

Param	Val	AR	FWE	AFD
Base		0.9344	0.0346	0.0035
M	50	0.9817	0.0269	0.0025
M	100	0.9690	0.0404	0.0038
M1	1	0.9442	0.0462	0.0231
M1	50	0.9404	0.0404	0.0009
M3	0	0.9371	0.0500	0.0049
M3	50	0.9413	0.0462	0.0050
Strength	0.05	0.0323	0.0135	0.0109
Strength	0.1	0.4577	0.0423	0.0092
Strength	0.25	1.0000	0.0365	0.0033
Correlation	0	0.9438	0.0404	0.0039
Correlation	0.1	0.9383	0.0365	0.0037
alpha	0.01	0.8675	0.0058	0.0006
Return.Generation	hsu	0.9406	0.0327	0.0032