Central European University

Master's Thesis

A STATISTICAL TEST TO CONTROL EXCESSIVE PARAMETER FITTING OF TRADING STRATEGIES

Author

Krisztián Szemán

Supervisor

Dr. Tamás Ferenci

This thesis is submitted in fulfillment of the requirements for the degree of Master of Science in Mathematics.

Mathematics department

Contents

Abstract. Checking the past performance of many variants of a trading strategy with different parameters and selecting the one with the highest observed Sharpe Ratio is a common, nevertheless ill-advised practice when optimizing trading strategies. While the portfolio with the highest Sharpe Ratio might be viewed as the best choice of a risk-averse investor (e.g. according to Modern Porfolio Theory), true Sharpe Ratios cannot be deducted from backtests. Observed Sharpe Ratios can only estimate them through an error. Among several similar trading strategies, having one with the highest observed Sharpe Ratio is inevitable, but whether it is due to randomness or actually enhanced performance needs investigation. In my thesis, I propose a hypothesis test that helps determine the answer to this question. This test is applied to a dataset containing the performance of 18 of iShares used by previous authors for two different timeframes. I conclude that the iShare with the highest realized Sharpe Ratio does not exhibit a significantly better performance than the average iShare in neither of those periods.

Introduction

This thesis proposes a statistical test for comparing realized Sharpe Ratios of many similar trading strategies. The goal of this test is to determine whether the highest of such realized Sharpe Ratios is significantly higher than their average.

The test is needed due to an ill-advised practice among those looking for profitable trading strategies for the purpose of monetary gains or publishing scientific papers. It is a common practice to check the historical performance of a strategy, and assume that the performance is going to be similar in the future. This is a reasonable assumption as long as we run one such test.

In reality, more often than not, this is not the case. For instance, strategies usually have adjustable parameters. Different parameter configurations will lead to different historical performance. One might naively choose the one that yields the best historical performance. Clearly, even if there is no underlying difference between the different configurations, the historical performance will differ due to randomness. Choosing the best one in this case leads to false inference about the performance, a performance that is likely to worsen in the future. With this in mind, it is essential to investigate whether the difference between the strategies is statistically significant. Since it is a standard method to choose the strategy with the best performance, it is natural to test whether that is significantly better than the average. The construction of such a test is the main result of my work.

The rest of my thesis is organized as follows: Section 1 contains some vital properties of the metric most commonly used for evaluating trading strategies, the Sharpe Ratio. Section 2 elaborates on the dangers of testing multiple trading strategies. Section 3 is a review of the literature concerning statistical methods to compare realized Sharpe Ratios. These methods are essential for the hypothesis test described in Section 4. The test is applied on empirical data in Section 5: the historical performance of 18 iShares (international stock indices) is tested. Lastly, Section 6 summarizes the thesis.

1. The Sharpe Ratio and its properties

In this section, I give an overview of some key charasteristics of the Sharpe Ratio, the most widely used metric in evaluating the performance of trading strategies and trading firms, and demonstrate a few properties that justify the usage of this metric. First, the realized Sharpe Ratio can be directly used to test the significance of excess returns. Secondly, under some general assumptions, the actual Sharpe Ratio is the ultimate metric to make decisions about asset allocation.

As mentioned above, trading strategies are often evaluated by their realized Sharpe Ratios. I aim to underline the importance of differentiating between this and the actual Sharpe Ratio (and point out the dangers of confusing them). The latter is almost always unknown in reality, and can only be estimated through an error. Throughout this section and the rest of my thesis, the former is referred to as realized, observed or estimated Sharpe Ratio, while the latter is called actual or true Sharpe Ratio.

1.1. Sharpe Ratio: Definition and typical values.

The Sharpe Ratio (denoted by SR), proposed by William F. Sharpe [13] is a straightforward and widely used metric for evaluating performance of asset managers, funds or even individual trading strategies. It is the expected value of the excess return of an invesment over some benchmark (e.g. risk-free bond or an appropriate stock index) divided by its standard deviation: if $(r_{p,1}, r_{p,2}, \dots, r_{p,T})$ is the return vector of an investment in a T-day period, and $(r_{f,1}, r_{f,2}, \dots, r_{f,T})$ is the risk-free return vector in that same period, then the Sharpe Ratio is

$$SR \doteq \frac{\mu}{\sigma},$$

where $\mu \doteq E\left(r_{p,t} - r_{f,t}\right)$ and $\sigma \doteq D\left(r_{p,t} - r_{f,t}\right)$ are the expected value and the standard deviation of excess returns, respectively. Note that for this formula to make sense, we have to assume that the excess returns have finite a second moment and are stationary and ergodic. These, however, are widely accepted assumptions across the literature. To avoid confusion, it is worth mentioning that in most usecases the annualized Sharpe Ratio is referred to, that is $SR_{ann} \doteq \frac{\mu}{\sigma} \cdot \sqrt{T}$, where $\sqrt{T}$ is a constant term included to represent the number of trading days in a year (T is typically between 250 and 256). While in the financial world, it is usually the annulized version that is referred to as the Sharpe Ratio, I will follow the convention in the literature in my thesis and denote the non-annualized version by SR.

The distinction is more important than it seems at first. This short paragraph does not aim to be the strictest possible in terms of terminology, but it does provide a feel for the usual range of Sharpe Ratios (see Table 1), and an example where the two versions

SR_{ann}	SR	Evaluation of fund/strategy	P(losing day)	P(losing year)
0	0	No value	50%	50%
0.5	0.032	Mediocre	48.7%	31%
1	0.063	Acceptable	47.5%	16%
2	0.127	Good	45%	2.3%
3	0.190	Excellent	42.5%	0.13%

TABLE 1. Some typical values of Sharpe Ratios. P stands for probability, and the probability of losing year is calculated assuming i.i.d., normally distributed daily returns (losing means performing worse than the benchmark).

of SR were confused. The table is included to emphasize that for practical purposes, one should consider annualized Sharpe Ratios from 0 to 3-4, that is, non-annualized Sharpe Ratios of 0 to roughly 0.2-0.25. With that said, Lo [10], for instance in Table 1 of their work evaluates "Asymptotic Standard Errors of Sharpe Ratio Estimators" while referring to the non-annualized version of the Sharpe Ratio. The author considers non-annualized Sharpe Ratios between 0.5 and 3, which is equivalent to annualized SRs of roughly 8 to 48. The standard errors (or any other attribute, for that matter) of estimators of Sharpe Ratios that high are clearly irrelevant from a practical point of view, and the table is almost surely the result of failing to make the distinction highlighted in this paragraph.

The remainder of this section is used to pinpoint some properties of the Sharpe Ratio.

1.2. **Sharpe Ratio and hypothesis testing.** The actual Sharpe Ratio of a strategy can not be observed directly, but it can be estimated from a time-series of excess returns. Let $(r_1, r_2, \ldots, r_T)$ be the observed excess returns of a portfolio over a T-day horizon, then the estimator of SR is $\hat{SR} \doteq \frac{\hat{\mu}}{\hat{\sigma}}$, where where $\hat{\mu} = \frac{\sum_{t=1}^{T}(r_t)}{T}$ is the mean of the observed daily returns, and $\hat{\sigma} = \sqrt{\frac{\sum_{t=1}^{T}\left(r_t - \frac{\sum_{s=1}^{T}(r_s)}{T}\right)^2}{T-1}}$ is their estimated standard deviation. A very convenient property of this realized Sharpe Ratio, $\hat{SR}$ is that it is a t-statistic up to a constant factor. Therefore, it can directly be used to test the significance of excess returns, will the null-hypothesis being $H_0 : E(r_t) = 0$.

Example. The excess returns of a portfolio exhibit a non-annualized Sharpe Ratio of 0.1 (annualized Sharpe Ratio of 1.58, assuming $T = 250$) over a 3-year horizon. Are the excess returns of this portfolio significantly greater than 0? We need to compute the p-value corresponding to the hypothesis test. Since $\hat{SR} \doteq \frac{\hat{\mu}}{\hat{\sigma}} = 0.1$, and the sample

size is $n = 750$, the corresponding t-value is $\frac{\hat{\mu}}{\frac{\hat{\sigma}}{\sqrt{n}}} = 0.1 \cdot \sqrt{n} = 2.738$. This translates to a p-value of 0.0015, as we are doing a one-sided t-test. This means the excess returns of the portfolio are significantly higher than 0 on the convential 5% level.

It is worth noting, however, that while in other scientific fields, e.g. medicine the convential 5% level is used, when testing the significance of excess returns, usually a threshold stricter than 5% is applied. Why is it justified? In his famous essay titled "Why Most Published Research Findings Are False", Ioannidis [5] defines PPV, the positive predictive value as the probability of a research finding being true, and gives the formula $PPV = \frac{(1-\beta) \cdot R}{(1-\beta) \cdot R + \alpha}$, where R is the ratio of the number of "true relationships" to "no relationships" among those tested in the field, and β is the type II error rate. The low signal-to-noise ratio in financial data and recent reproductions of financial studies suggest that R is relatively small in this area of research compared to others, so in order to achieve a similar PPV value, α should be lower (typical values might be 0.1% or 1%).

1.3. **Sharpe Ratio and asset allocation.** The Capital Asset Pricing Model (CAPM) is a widely used model helping security pricing in the financial world. One of the assumptions of CAPM is that the expected value and covariance of future asset returns is known. Using this, the risk-return profile of optimal portfolios for an investor lies on a curve called the Efficient Frontier. A portfolio is on the Efficient Frontier if and only if there is no portfolio that provides higher expected returns while having lower or equal standard deviation. Any possible portfolio that is not on it is sub-optimal, because there is an opportunity to achieve higher expected returns with the same level of risk, or the same return with lower risk. Among optimal portfolios, there is a trade-off between risk and return. If an investor wants to have higher expected return, they have to accept a higher risk. This trade-off is not constant. As the expected returns increase, the investor has to endure proportionally increasing additional risk to have even higher expected returns.

In Modern Portfolio Theory (MPT), it is possible to invest in the risk-free asset, a security with a guaranteed return known in advance (the closest thing to it on the market is usually considered to be the 3-month US Treasury Bill). In this case, the optimal asset allocation possibilities of an investor lie on a line called the Capital Market Line (CML). Any portfolio on the CML is the combination of the risk-free asset and one element of the Efficient Frontier (this is illustrated on Figure 1.1). It is easy to see that this element is exactly the one with the highest expected excess return to standard

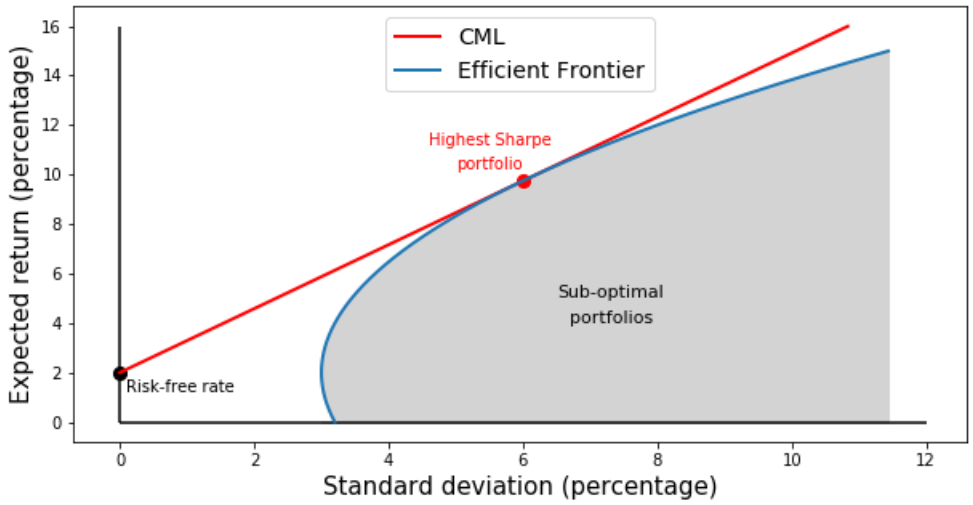

FIGURE 1.1. Illustration of the Efficient Frontier and the Capital Market line

deviation ratio, i. e. the highest Sharpe Ratio. For portfolios on the CML, in contrast to the ones on the Efficient Frontier, the trade-off between risk and return is constant.

Neither CAPM nor MPT are perfect reflections of the real world, as many of their assumptions do not hold in reality. Most notably, future expected values and covariances of asset returns are unknown. Nevertheless, the most important objective for portfolios, strategies or trading firms is often to have the highest Sharpe Ratio possible.

1.4. **Estimator of the Sharpe Ratio.** As mentioned previously, based on the historical performance of a fund or a trading strategy, we can only calculate the realized Sharpe Ratio, $\hat{SR}$,which is an estimator of the actual Sharpe Ratio, and this estimator has a variance. This variance is crucial to discuss, as it is important in works mentioned later in my thesis considering the comparison of Sharpe Ratios of two or more trading strategies. Using CLT and the delta method, Lo [10] derives that under the assumption of i.i.d. (independent, identically distributed) returns following a normal distribution,[1] $\sqrt{T}\cdot\left(\hat{SR}-SR\right) \overset{a}{\sim} N\left(0, 1+\frac{SR^2}{2}\right)$, where T is the length of the period in days. Mertens [12] extends this work, showing that the correct formula for i.i.d. returns with a finite fourth moment is $\sqrt{T}\cdot\left(\hat{SR}-SR\right) \overset{a}{\sim} N\left(0, 1-\gamma_3\cdot SR+\frac{\gamma_4-1}{4}\cdot SR^2\right)$, where γ_3 and γ_4 are the third and fourth moments of excess returns, respectively.

[1]The author mistakenly only requires the existence of a finite second moment. However, Mertens [12] shows that the formula only holds for i.i.d normal returns, while giving a correct formula for the more general case.

We can see that negative skewness and excess kurtosis can inflate the variance of the estimator, and it is worth noting that the variance given by $1 - \gamma_3 \cdot SR + \frac{\gamma_4 - 1}{4} \cdot SR^2$ is bounded by $\left(1 - \frac{1}{2}\gamma_3 SR\right)^2$ from below due to the Pearson's inequality: $\gamma_4 \geq 1 + \gamma_3^2$. Moreover, its worth noting that for typical values of SR, the quadratic term is fairly small, so given a symmetric return distribution, the aforementioned variance terms are close to one. However, negative skewness of returns is common for hedge funds and most other market participants due to long conjencture periods and short, severe drawdowns. As the formula suggests, if the distribution of returns exhibits a negative skewness, the variance of the estimator can be significantly higher.

2. Backtesting, fitting and overfitting trading strategies

As Bailey and Lopez De Prado [1] discusses, modern trading strategies are often built on patterns found in financial data that can be exploited in a systematic way. Given the historical data, it is possible to check how such a strategy would have performed in the past (in-sample). This procedure is called backtesting. The historical performance of a backtested strategy is often evaluated based on its realized Sharpe Ratio. As mentioned earlier in my thesis, it is also straightforward to test the significance of the excess returns based on this metric.

However, when looking for profitable trading strategies nowadays, an individual is highly unlikely to test only one. With the development in high-performance computing over the last decade, it has become possible to backtest millions of trading strategies in a fairly short time. While there are undeniable advantages of this, one great drawback is that the probability of false discovery increases tremendously. On a chosen significance level α, the probability of a false discovery assuming no effect (in this case, seemingly significant positive excess returns that are zero in reality) is α. In this case, for a million tests, we would get in expectation α times one million false positive results (fifty-thousand for $\alpha = 5\%$, a thousand for $\alpha = 0.1\%$). Even without assuming independence of the tests, we are likely to have an abundance of false positive results.

Obviously, labelling false discoveries as profitable strategies will lead to disappointing future (out-of-sample) performance. This inconsistency of in-sample and out-of-sample performance is often called overfitting, the term adopted from machine learning to describe a similar phenomenon.

A trading strategy often has customizable parameters. For example, consider one of the early works relying on backtesting by Jegadeesh and Titman [6]. The authors demonstrated the momentum effect on the stock market. Momentum means stocks that performed well in the recent past are likely to continue to do so, while those that performed poorly also continue to do so more often than not. This means taking long positions in the former, and short positions in the latter stocks result in excess returns for an investor. In their work, they examined past performance of the latest 3, 6, 9, and 12 months, with holding periods of 3, 6, 9, and 12 months (16 possibilities). They also considered a second set of 16 strategies that skip a week between the portfolio formation period and the holding period. There were also three separate cases in terms of long-short position : long only, short only, and long-short portfolios. Altogether this already means 96 different strategies. In the paper, the backtested results of all the

strategies are published. With that said, there are still some other arbitrary choices that were made by the researchers, e.g.:

- the way to sort stocks: putting them into deciles based on their returns in the past period (long positions taken in the best, short in the worst decile)
- the set of examined stocks

The key thing to note is that even such a simple idea as momentum can have several customizable parameters, and different parameter combinations obviously lead to different backtested performance. It is a natural idea to look for the parameter combination that yields the best historical performance. Historical performance is usually evaluated by the observed SR ($\hat{SR}$), and a high SR is desirable due to reasons discussed before. Since the realized Sharpe Ratio is not the true one, and it is subject to some estimation error, carelessly selecting the strategy with the highest $\hat{SR}$ can result in overfitting, meaning to out-of-sample performance of a strategy can be much worse than its performance in-sample. The following example shows how equally powerful strategies can have very different realized Sharpe Ratios.

Example. To mimic the behaviour of trading strategies, I simulated 100 return processes for an 1000-day horizon (4 years). In this example, every strategy has the same true Sharpe Ratio of 0.1 (1.58 annualized), the daily returns follow a multivariate normal distribution, $\rho(r_{i,t}, r_{j,s}) = 0.3$ if $t = s$ and $i \neq j$, and $\rho(r_{i,t}, r_{j,s}) = 0$ if $t \neq s$. This means returns on different days are independent from each other, but returns of different strategies correlate slightly on the same day (to capture the effect of backtesting somewhat similar strategies). The objective of Figure 2.1 is to show the effect of randomness on the performance and realized Sharpe Ratio of strategies, and emphasize how ill-advised it is to choose to best one without further consideration.

While all the strategies have equal Sharpe Ratios, hence in expectation, they would perform equally well in the future, the highest observed SR in this example is a superb 0.176 (and in general, for this simulation the expected value of the maximum realized SR is 0.17), which is almost double the value of the true SR, 0.1. Investing in this strategy would likely lead to a drop of around 40-45% out-of-sample compared to in-sample.

Generalizing the example above, Table 2 shows the expected value of the best realized SR for various numbers of backtested strategies and pairwise correlations (with the time horizon and true SR remaining remains 1000 days and 0.1, respectively):

Not surprisingly, the degree of overestimation of the true Sharpe Ratio increases as the number of backtested strategies increase. Besides that, overestimation is lower if

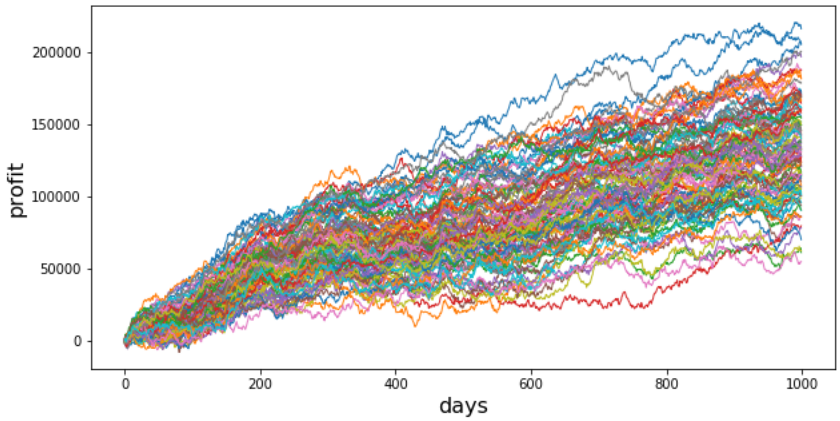

FIGURE 2.1. Modeling the cumulative profit of 100 correlated strategies with equal true Sharpe Ratios through 1000 days

	10	100	1000
0%	0.149	0.180	0.204
25%	0.151	0.177	0.198
50%	0.134	0.157	0.175
75%	0.133	0.148	0.160

TABLE 2. Expected value of the maximum of observed Sharpe Ratios given 10, 100 or 1000 toy strategies with pairwise correlation of 0, 25, 50 and 75%. The values are estimations arising from 100 thousand simulations using Python for each case

correlation between the backtested strategies is higher. That means trying out strategies that are actually very similar mitigates the extent of overfitting.

To further worsen things, this harmful effect can be enhanced if one chooses to parametrize the best strategy in some new way. They may create and backtest many similar variants of it, in order to further improve in-sample performance. Due to randomness, this is likely to lead to an even higher maximum of in-sample Sharpe Ratios, and out-of-sample performance that is even more inconsistent with it.

Why is this a problem? One could argue that even though the backtest is not realistic for the strategy with the highest realized SR, since all strategy have the same true SR, it does not matter which one is chosen (in terms of expected future performance). So we might as well choose the one with the best backtested performance. While this is true in itself, backtests are much more valuable when they are realistic compared to when they are not. One of the reasons is it makes it easier to combine strategies in a way that achieves the optimal portfolio. As mentioned previously, Modern Portfolio

Theory's assumption of known expected values and covariance of future returns is not true. But if it were, one would have the opportunity to build the best possible portfolio. The better estimation (the more realistic backtest) one has, the closer they get to that goal.

The purpose of the example above was to show the dangers of senseless parameter fitting of a trading strategy. However, this is not to say that any kind of parameter fitting is meaningless. There are certainly circumstances when it is reasonable to try it and when it actually adds value. The main aim of my thesis is to help identify these situations. The hypothesis test in Section 4 asks the following question: given N backtested strategies, does the seemingly optimal strategy (one with the highest observed Sharpe Ratio) have a higher true Sharpe Ratio than the average Sharpe Ratio of all the strategies? In other words, are we right to claim that one strategy is better than the rest, or is its outperformance only due to randomness?

3. Comparing realized Sharpe Ratios

The main aim of the previous two sections was to pinpoint the difference between actual and realized Sharpe Ratios. The most important difference from this work's perspective is when it comes to the comparison of Sharpe Ratios. While between different actual Sharpe Ratios over the same time horizon, it is correct to say that the higher is the better, it is far from true for realized Sharpe Ratios, as their difference could be simply due to randomness. As the hypothesis test described in the next section heavily relies on methods for comparing multiple realized Sharpe Ratios, I describe important previous results published in this area over the next few paragraphs.

First, Jobson and Korkie [7] established a test for the equality of multiple realized Sharpe Ratios, and their formula was later corrected by Memmel [11]. However, the authors in both cases relied on the assumption of i.i.d. returns following a multivariate normal distribution. Probability distributions in financial time series often exhibit non-zero skewness, as well as tails that are heavier than that of the normal distribution, and they are also serially correlated.

As Ledoit and Wolf [8] shows, the Jobson-Korkie test and its corrected version are both invalid when daily returns exhibit tails heavier than the normal distribution or are of time series nature. They propose a more robust test that does not require neither serial indepence nor normally distributed daily returns. However, the paper only considers comparing two realized Sharpe Ratios. Since part of their methodology is important for the rest of my work, I discuss it in more detail below.

Let $(r_{1,1}, r_{1,2}, \ldots, r_{1,T})$ and $(r_{2,1}, r_{2,2}, \ldots, r_{2,T})$ be the observed excess returns of two portfolios over a T-day horizon. The two return series are only assumed to constitute a strictly stationary time series. This means that the bivariate distribution of returns is unchanged over time. The mean of this bivariate distribution is $\mu = (\mu_1, \mu_2)$ and its covariance is $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$. Hence, the Sharpe Ratios are given by: $SR_1 = \frac{\mu_1}{\sigma_1}$ and $SR_2 = \frac{\mu_2}{\sigma_2}$. The sample means and sample variances of the returns observed are denoted $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2)$ and $\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{bmatrix}$, respectively. The goal is to test the significance of the difference of two realized Sharpe Ratios, that is $\hat{\Delta} \doteq \hat{SR_1} - \hat{SR_2} = \frac{\hat{\mu}_1}{\hat{\sigma}_1} - \frac{\hat{\mu}_2}{\hat{\sigma}_2}$. Similarly to any other hypothesis test, one needs to derive the distribution of the test statistic, $\hat{\Delta}$ under the assumption that the null-hypothesis holds, that is, $\Delta \doteq SR_1 - SR_2 = 0$.

The distribution of $\hat{\Delta}$ can be derived using the multivariate version of the delta method.

Theorem. *Univariate delta method*

Let $(\Omega, \mathscr{F}, \mathbf{P})$ *be a probability space, and* (X_n) *a sequence of real-valued random variables:* $X_n : \Omega \to \mathbb{R}$, $\forall n \epsilon \mathbb{N}$. *Assume that* (X_n) *satisfies* $\sqrt{n}(X_n - x) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ *for some* $x, \sigma \epsilon \mathbb{R}$, *and suppose* $g : \mathbb{R} \to \mathbb{R}$ *is a function for which* $g'(x)$ *exists and* $g'(x) \neq 0$. *Then* $\sqrt{n}(g(X_n) - g(x)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 g'(x)^2\right)$.

Proof. The proof is basically an application of the Taylor's theorem. The sketch of it is as follows: the Taylor expansion of $g(X_n)$ is $g(X_n) = g(x) + g'(x)(X_n - x) + \varepsilon$, therefore $\sqrt{n}(g(X_n) - g(x)) = \sqrt{n}(g'(x)(X_n - x)) + \sqrt{n}\varepsilon$. On the right-hand side, $\sqrt{n}(g'(x)(X_n - x))$ is a linear transformation of $(X_n - x)$, thus $\sqrt{n}(g'(x)(X_n - x)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 g'(x)^2\right)$, and $\sqrt{n}\varepsilon$ converges to 0 in probability as $X_n \to x$ in probability. Applying Slutsky's theorem, $\sqrt{n}(g(X_n) - g(x)) = \sqrt{n}(g'(x)(X_n - x)) + \sqrt{n}\varepsilon \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 g'(x)^2\right)$. □

The multivariate delta method is not much more than a straightforward extension of the univariate version.

Theorem. *Multivariate delta method*

Let $(\Omega, \mathscr{F}, \mathbf{P})$ *be a probability space, and* (X_n) *a sequence of k-dimensional real-valued random variables:* $X_n : \Omega \to \mathbb{R}^{\mathrm{k}}$, $\forall n \epsilon \mathbb{N}$. *Assume there exists* $x = (x_1, x_2, \ldots x_k) \epsilon \mathbb{R}^{\mathrm{k}}$ *such that* $\sqrt{n}(X_n - x) \xrightarrow{d} N(0, \Sigma)$, *where* $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \ldots & \sigma_k^2 \end{bmatrix}$ *is a symmetric positive semi-definite covariance matrix. Let* $g : \mathbb{R}^{\mathrm{k}} \to \mathbb{R}$ *be a function with continous partial derivatives* $g_1', g_2', \ldots g_k'$, *and assume that* $\sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_{ij} g_i'(x_i) g_j'(x_j) > 0$. *Then* $\sqrt{n}(X_n - x) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(x)^T \Sigma \nabla g(x)\right)$.

The proof is similar to the univariate case so it is omitted.

When comparing realized Sharpe Ratios, let $u = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, $\hat{u} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ and $f : \mathbb{R}^4 \to \mathbb{R}$ a function such that $f(a, b, c, d) = \frac{a}{\sqrt{c}} - \frac{b}{\sqrt{d}}$. This means $\Delta \doteq f(u)$ and $\hat{\Delta} \doteq f(\hat{u})$.

By the Central Limit theorem, we have that $\sqrt{n}(\hat{u} - u) \xrightarrow{d} \mathcal{N}(0, \Omega)$. Previous authors, like Jobson and Korkie [7], Memmel [11] assume $\Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 2\sigma_1^4 & 2\sigma_{12}^2 \\ 0 & 0 & 2\sigma_{12}^2 & 2\sigma_2^4 \end{bmatrix}$,

but that only holds for returns following a multivariate normal distribution without serial dependence. In contrast, Ledoit and Wolf [8] does not specify Ω any further at this point.

They proceed by applying the multivariate delta method to $\hat{u}$, u and f. This yields $\sqrt{n}\left(\hat{\Delta}-\Delta\right)=\sqrt{n}\left(f\left(\hat{u}\right)-f\left(u\right)\right)\xrightarrow{d}N\left(0,\nabla f\left(u\right)^{T}\Omega\nabla f\left(u\right)\right)$, with $\nabla f\left(u\right)^{T}=\left(\frac{1}{\sqrt{c}},-\frac{1}{\sqrt{d}},-\frac{a}{2\sqrt{c^{3}}},\frac{b}{2\sqrt{d^{3}}}\right)$. Once this limiting variance is known, it is straightforward to construct the hypothesis test and assign a p-value to it: assuming $H_0:\Delta=0$, $\sqrt{n}\hat{\Delta}$ follows a normal distribution with 0 mean and $\theta\doteq\nabla f\left(u\right)^{T}\Omega\nabla f\left(u\right)$ variance. Therefore, the p-value is given by $p=2\left(1-\Phi\left(\frac{\sqrt{n}\left|\hat{\Delta}\right|}{\theta}\right)\right)$.

The key result to note here is that $\hat{\Delta}$ follows an asymptotically normal distribution with Δ mean. Ledoit and Wolf further proposes multiple ways to estimate Ω in a manner that is consistent with heteroskedasticity and autocorrelation of the returns. My thesis does not include the description of these methodologies, as the hypothesis test in the next section uses a simpler methodology that could estimate the variance of $\sqrt{n}\left(\hat{\Delta}-\Delta\right)$ directly.

Extending the previous result, Wright [14] proposes a test for the equality of multiple realized Sharpe Ratios. He also concludes that the methodology proposed by Leung and Wong [9] in their test is inappropriate. The previous three papers all tested the equality of the observed Sharpe Ratios of 18 iShares for the period between 1996 and 2003. While Leung and Wong found the performance of the 18 iShares to be distinguishable, both Ledoit and Wolf and Wright contradicts this result, deeming the 18 realized Sharpe Ratios not to be statistically significantly different. In Section 5 of my thesis, a similar test is carried out to determine whether the performance of the best-performing iShare is significantly better than the rest.

4. A statistical test to justify (or reject) parameter fitting of trading strategies

In Section 2, I emphasized the dangers of comparing realized Sharpe Ratios based purely on their values. In Section 3, I described a more appropriate way to compare realized Sharpe Ratios. These lay the foundation to pose and answer the question that is the focal point of this thesis: given N backtested trading strategies, is the performance of the best one significantly different than the average?

We are given historical, backtested performance of N strategies over a T-day horizon. In this setting, it is best to imagine these strategies as different variants of the same idea (e.g. momentum), with different parameters. Let $r_{n,t}$ denote the excess return of the n-th strategy over some universal benchmark for $t = 1, 2, \ldots, T$ and $n = 1, 2, \ldots, N$. These excess returns are required to be stationary and ergodic and to have finite fourth moments. The following notations are consistent with the previous sections: $\mu_n \doteq E(r_{n,t})$ is the expected daily return of the n-th strategy, $\sigma \doteq D(r_{n,t})$ is its standard deviation, and $\hat{\mu}_n$ and $\hat{\sigma}_n$ are their sample estimates. $SR_n \doteq \frac{\mu_n}{\sigma_n}$ is the true Sharpe Ratio of the n-th strategy, while $\hat{SR}_n \doteq \frac{\hat{\mu}_n}{\hat{\sigma}_n}$ is the corresponding observed Sharpe Ratio.

The null-hypothesis is that none of these strategies is truly superior to the others, so parameter fitting (looking for the parameter combination with the best historical performance) is useless. This means $H_0 : \max_k \left(SR_k - \frac{1}{N} \sum_{i=1}^{N} SR_n \right) = 0$. Note that while this is equivalent to assuming $SR_1 = SR_2 = \ldots = SR_N$, the aim here is not to test the equality of all the Sharpe Ratios. Let κ denote the difference of the maximum of observed Sharpe Ratios and their average: $\kappa \doteq \max_k \left(\hat{SR}_k - \frac{1}{N} \sum_{i=1}^{N} \hat{SR}_n \right)$.

In order to make the statistical test work, we need to derive the distribution of κ under the null hypothesis. Denoting $M \doteq \begin{bmatrix} 1-\frac{1}{N} & -\frac{1}{N} & \ldots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & \ldots & -\frac{1}{N} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{N} & -\frac{1}{N} & \ldots & 1-\frac{1}{N} \end{bmatrix}$, κ is the maximal element of the vector $\left(\hat{SR}_1 - \frac{1}{N} \sum_{j=1}^{N} \hat{SR}_j, \ldots, \hat{SR}_N - \frac{1}{N} \sum_{j=1}^{N} \hat{SR}_j \right)^T = M \left(\hat{SR}_1, \ldots, \hat{SR}_N \right)^T$, so the first task is to derive the distribution of $\left(\hat{SR}_1, \ldots, \hat{SR}_N \right)^T$. The methodology I propose to do this is very similar to that of Ledoit and Wolf [8]. Let $u = (\mu_1, \mu_2, \ldots, \mu_N, \sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$, $\hat{u} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_N, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_N^2)$ and $f : \mathbb{R}^{2N} \to \mathbb{R}^N$ such that $f(a_1, a_2, \ldots, a_N, b_1, b_2, \ldots, b_N) = \left(\frac{a_1}{\sqrt{b_1}}, \frac{a_2}{\sqrt{b_2}}, \ldots, \frac{a_N}{\sqrt{b_N}} \right)$. Hence

$(SR_1, SR_2, \ldots, SR_N) = f(u)$ and $\left(\hat{SR}_1, \hat{SR}_2, \ldots, \hat{SR}_N\right) = f(\hat{u})$. I assume $\sqrt{T} \cdot (\hat{u} - u) \xrightarrow{d} \mathcal{N}(0, \Omega)$. This follows from the Central Limit Theorem under mild regularity conditions. According to Ledoit and Wolf, "it is sufficient to have finite $4 + \delta$ moments, where δ is some small positive constant, together with an appropriate mixing condition". Given this, the multivariate delta method implies that

$$\sqrt{T} \cdot \left(\left(\hat{SR}_1, \hat{SR}_2, \ldots, \hat{SR}_N\right) - (SR_1, SR_2, \ldots, SR_N)\right) =$$

$$= \sqrt{T} \cdot (f(\hat{u}) - f(u)) \xrightarrow{d} \mathcal{N}\left(0, \nabla f(u)^T \Omega \nabla f(u)\right),$$

where $\nabla f(a_1, a_2, \ldots, a_N, b_1, b_2, \ldots, b_N) = \left(\frac{1}{b_1^{0.5}}, \frac{1}{b_2^{0.5}}, \ldots, \frac{1}{b_N^{0.5}}, \frac{a_1}{b_1^{1.5}}, \frac{a_2}{b_2^{1.5}}, \ldots, \frac{a_N}{b_N^{1.5}}\right)$. This means $\sqrt{T}\left(\left(\hat{SR}_1, \hat{SR}_2, \ldots, \hat{SR}_N\right) - (SR_1, SR_2, \ldots, SR_N)\right)$ follows an asymptotically normal distribution with zero mean and covariance matrix $\Psi_0 = \nabla f(u)^T \Omega \nabla f(u)$.

Using the result above leads to

$$\sqrt{T}\left(\hat{SR}_1 - \frac{1}{N}\sum_{j=1}^{N}\hat{SR}_j, \hat{SR}_2 - \frac{1}{N}\sum_{j=1}^{N}\hat{SR}_j, \ldots, \hat{SR}_N - \frac{1}{N}\sum_{j=1}^{N}\hat{SR}_j\right)^T =$$

$$= \sqrt{T} M \left(\hat{SR}_1, \hat{SR}_2, \ldots, \hat{SR}_N\right)^T \xrightarrow{d} \mathcal{N}(M(SR_1, SR_2, \ldots, SR_N), M\Psi_0 M).$$

To simplify things, $M(SR_1, SR_2, \ldots, SR_N) = 0$ under the null hypothesis, therefore

$$\sqrt{T}\left(\hat{SR}_1 - \frac{1}{N}\sum_{j=1}^{N}\hat{SR}_j, \ldots, \hat{SR}_N - \frac{1}{N}\sum_{j=1}^{N}\hat{SR}_j\right) \xrightarrow{d} \mathcal{N}(0, M\Psi_0 M).$$

As κ is the maximal element of the vector on the left-hand side, it can be viewed as the maximum of a multivariate normal distribution with zero mean and $\Psi \doteq M\Psi_0 M$ covariance.

At this point, I diverge from Ledoit and Wolf's methodology in estimating Ψ. Instead of estimating Ω first, and then using $\Psi = M\left(\nabla f(u)^T \Omega \nabla f(u)\right) M$, I propose $\frac{\Psi_0}{\sqrt{T}}$ to be approximated directly, as it is basically the covariance of the N realized Sharpe Ratios. The j-th element of the i-th row of this matrix, $\frac{\Psi_0}{\sqrt{T}}$ is $\text{cov}\left(\hat{SR}_i, \hat{SR}_j\right)$. Using the assumption that returns are stationary and ergodic, the time series of T days can be divided into B blocks, each having a length of $\tau \doteq \left[\frac{T}{b}\right]$, where $[\cdot]$ denotes the integer part of a real number. Let $\hat{SR}_{n,b}$ denote the realized Sharpe Ratio of the n-th strategy over the b-th block for every $n \epsilon \{1, \ldots, N\}$ and $b \epsilon \{1, \ldots, B\}$. That means $\hat{SR}_{n,b}$ is the average of the set of returns $\{r_{b\tau}, r_{b\tau+1}, \ldots, r_{b\tau+\tau-1}\}$ divided by its estimated standard deviation.

Since $\sqrt{\tau}\left(\left(\hat{SR}_{1,b}, \hat{SR}_{2,b}, \ldots, \hat{SR}_{N,b}\right) - \left(SR_{1,b}, SR_{2,b}, \ldots, SR_{N,b}\right)\right) \xrightarrow{d} N\left(0, \Psi_0\right)$, we can approximate $\frac{\Psi_0}{\sqrt{\tau}}$ from the sample of Nb realized Sharpe Ratios in the usual way to estimate covariance matrices.

Finally, κ is the maximum of a normal distribution with 0 mean and a covariance matrix that is estimated in the way described above. As much as the distribution of such a random variable is important for the statistical test to work, it is also not easy to derive it. This topic itself is an area of research as well. For instance Botev et. al. [4] describes a way to estimate the tail distribution of the maximum of non-independent Gaussian random variables. An explicitly formula is also possible to give, it is however extremely inconvenient to work with.

To illustrate one way of deriving it, I include the case of two variables with variance of 1. For two jointly Gaussian variables ξ_1, ξ_2 with zero mean, unit variance and ρ correlation, the cumulative distribution function of $\max\left(\xi_1, \xi_2\right)$ is $F\left(x\right) = 2\int\limits_{-\infty}^{x} \phi\left(z\right)\Phi\left(\frac{1-\rho}{\sqrt{1-\rho^2}}z\right)dz = 2\int\limits_{-\infty}^{x}\phi\left(z\right)\int\limits_{-\infty}^{\frac{1-\rho}{\sqrt{1-\rho^2}}z}\phi\left(w\right)dwdz$. For more than two variables, this expression could be extended by one more integral for every variable, and the upper limits of each integral could be calculated from the Cholesky-decomposition of the correlation matrix.

Since this part is certainly not the focal point of my thesis, I settle for computer simulations to estimate this cumulative distribution function and hence a p-value. In particular, using the multivariate.normal function of the Python package scipy, I simulate N jointly Gaussian random variables, $\xi_1, \xi_2, \ldots, \xi_N$ with zero mean and $\frac{\hat{\Psi}_0}{\sqrt{T}}$ covariance. Let $V \doteq \max\limits_{i\epsilon\{1,\ldots,N\}}\xi_i - \frac{1}{N}\sum\limits_{j=1}^{N}\xi_j$. I repeat this K times, denoting the corresponding values of the difference between the maximum and the average by $V_1, V_2, \ldots, V_K$. The estimator of the p-value is then derived as $\frac{1}{K}\sum\limits_{i=1}^{K}\mathbf{1}_{V_i>\kappa}$, where $\kappa = \max\limits_{k}\hat{SR}_k - \frac{1}{N}\sum\limits_{i=1}^{N}\hat{SR}_n$ and $\mathbf{1}$ denotes the indicator function.

To finish this section, before applying the test described to empirical data, it is important to note that there is an existing alternative to tackle the problem that I described in my thesis. Using the concept of the Probabilistic Sharpe Ratio (detailed by Bailey and Lopez De Prado [3]), Bailey and Lopez De Prado proposes the Deflated Sharpe Ratio [2]. The Deflated Sharpe Ratio, as its name suggests, aims to deflate the highest observed Sharpe Ratio given N backtested strategies. However, their formula in equation (2) on page 8 is incorrect. The full context can be found in the paper

referenced, as here I resort to merely pointing out the error. The formula is as follows: $D\hat{S}R = P\hat{S}R\left(\hat{SR}_0\right) = Z\left(\frac{\left(\hat{SR}-\hat{SR}_0\right)\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\hat{SR}+\frac{\hat{\gamma}_4-1}{4}\hat{SR}^2}}\right)$. According to the authors, under the null-hypothesis $\hat{SR} = \hat{SR}_0$, $\frac{\left(\hat{SR}-\hat{SR}_0\right)\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\hat{SR}+\frac{\hat{\gamma}_4-1}{4}\hat{SR}^2}}$ follows a normal distribution ($\hat{SR}$ is the maximum observed Sharpe Ratio across N strategies). However, using Mertens' formula, the standard error of $\left(\hat{SR}-\hat{SR}_0\right)\sqrt{T-1}$ is not $\sqrt{1-\hat{\gamma}_3\hat{SR}+\frac{\hat{\gamma}_4-1}{4}\hat{SR}^2}$, but $\sqrt{1-\gamma_3 SR+\frac{\gamma_4-1}{4}SR^2}$. However, SR in this case is unknown, and $\hat{SR}$ is surely not a consistent estimator of it, as it tends to overestimate it. Furthermore, the question here is whether $\hat{SR}$ is significantly bigger than $\hat{SR}_0$. Therefore it is incorrect to use $\hat{SR}$ instead of SR in the denominator of the formula above.

5. Testing on empirical data: the case of iShares

This section details the results of the hypothesis test introduced in the previous section, applied on the iShares data. The beginning is a description of the empirical data the test is applied to.

Since March 1996, 17 exchange-traded funds (ETFs) are traded on the American Stock Exchange. They are used to track the Morgan Stanley Capital International (MSCI) foreign stock market indices. Each of these ETFs is supposed to track the stock market index of a single country, e.g. Germany, France or Singapore. These ETFs were first known as iShares. Leung [9], Ledoit and Wolf [8] and Wright [14] all analyzed these iShares, with SPY, the ETF tracking the index of the US stock market being treated as the eighteenth iShare. Despite the fact that eventually, many more iShares began trading (as of 2020, there are over 200 of them), in accordance with previous authors, I used these same 18 iShares they did.

Passive investors often buy a collection of stocks or indices and hold it for a longer period of time. One such investor might want to figure out which one is the best to buy and hold out of the 18 iShares? These are 18 very similar trading strategies (the performance of iShares are highly correlated indeed, as shown below), with the one adjustable parameter being the country chosen. They have different realized Sharpe Ratios, but it is unclear whether that difference is statistically significant. The hypothesis test of the previous section is an appropriate tool to figure it out.

Before introducing my results, I lay out the those of previous authors. Comparing the 18 iShares, Leung [9] found that their performance is distinguishable. However, Ledoit and Wolf [8] contradicted this based on their pairwise test. They declared that the realized Sharpe Ratios of any two iShares is not statistically different. Wright [14] pointed out the mistake in Leung's approach, and his test for the equality of multiple Sharpe Ratios confirmed the result of Ledoit and Wolf. The examined time period in all three cases was a period of 1961 trading days between 1996 and 2003.

In my thesis, I ask a question that is different from the equality of all Sharpe Ratios. I investigate whether the performance of the historically best iShare is significantly different from the average performance of them. For the period between 1996 and 2003, the presumed answer based on the literature is likely that it is not. In addition to that, the period between 2004 and the beginning of 2020 is also considered in a separate test.

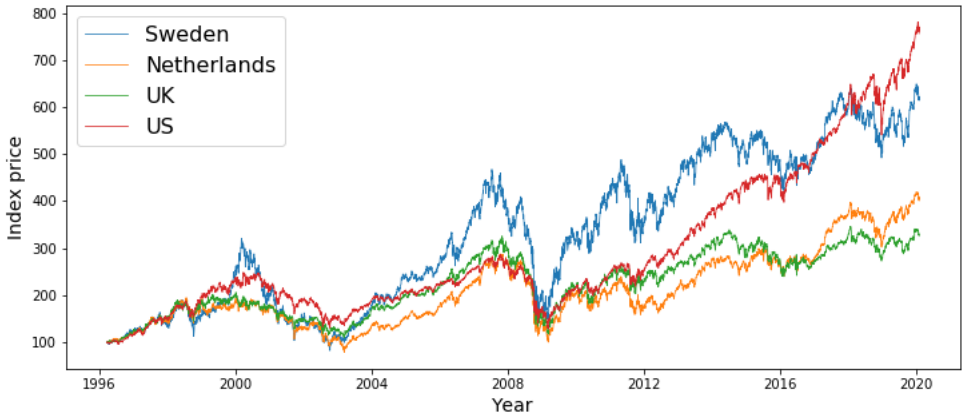

FIGURE 5.1. Normalized price of the 4 iShares: Sweden, the Netherlans, the UK and the US between 1996 and 2020 (1996=100)

Country	1996-2003	2004-2020
Australia	0.277	0.287
Austria	0.226	0.179
Belgium	0.226	0.241
Canada	0.423	0.294
France	0.384	0.210
Germany	0.185	0.227
Hong Kong	-0.007	0.331
Italy	0.447	0.041
Japan	-0.162	0.185
Malaysia	-0.169	0.272
Mexico	0.257	0.288
Netherlands	0.151	0.287
Singapore	-0.204	0.328
Spain	0.498	0.145
Sweden	0.229	0.261
Switzerland	0.193	0.423
United Kingdom	0.294	0.178
United States	0.396	0.479

TABLE 3. Realized Sharpe Ratios of all iShares in the period 1996-2003 and 2004-2020 (annualized)

Figure 5.1 shows the performance of 4 out of the 18 iShares between 1996 and 2020. Table 3 contains the realized Sharpe Ratios of all iShares in the period 1996-2003 and 2004-2020. The values are annualized for easier interpretation.

block size	EWD (1996-2003)	SPY (2004-2020)
30	67.1	58.3
60	67.2	53.5
90	66.8	51.8
120	67.4	49.8

TABLE 4. p-values for the two hypothesis tests using different block sizes for estimating Ψ_0

Between 1996 and 2003, the best performing ETF was the EWD, the index tracking the Spanish stock market, while in the second period from 2004 to 2020 it was the SPY, the ETF tracking S&P 500, the index of the US stock market. Their annualized observed Sharpe Ratios were 0.498 and 0.479, respectively. The daily returns of different ETFs were also very highly correlated: in the first period, the observed average pairwise correlation was 39%, while in the second one it rose to 75%. This increase can be explained partly by the continuous globalization and markets becoming more and more connected. One other reason is that while the first period does not exhibit any particular economic recession, the second period includes the financial crisis in 2008, and the first few days of markets reacting to the 2020 Coronavirus crisis: during such periods, correlation between asset returns grows.

Since in my thesis, I have not yet established any rule of thumb for the recommended number of blocks and the block sizes, the covariance matrix of realized Sharpe Ratios was approximated with 4 different block sizes, using $\tau = 30, 60, 90$ and 120 days. The two periods contained 1961 and 4048 trading days, respectively. Table 4 contains the p-values answering the following two questions:

- did EWD significantly outperform the average iShare in the period between 1996 and 2003?
- did SPY significantly outperform the average iShare in the period between 2003 and 2020?

The table shows that the outperformance of both EWD and SPY are only due to randomness. The p-values are fairly robust with respect to the block size in the first period, and a bit less so in the second. Still, in both cases the p-values are clearly statistically insignificant for all of the block sizes.

The results for the first period are absolutely not surprising, given that Ledoit and Wolf and Wright already reached similar conclusions for slightly different questions. On the other hand, the United States is known for having the strongest economy in the world, so the performance of SPY not being significantly better than the average

between 2004 and 2020 might be surprising for many. However, the realized (non-annualized) Sharpe Ratio of SPY is only 0.0303 for the examined period. Using Mertens' formula, given that the returns are negatively skewed and exhibit large kurtosis, with $\gamma_3 = -0.163$ and $\gamma_4 = 15.50$, the standard error of this estimator is $\frac{1+0.163 \cdot SR + 3.625 \cdot SR^2}{\sqrt{4048}} > 0.0157$ assuming $SR > 0$, where SR is the true Sharpe Ratio of SPY. The average iShare has a realized Sharpe Ratio of 0.0169, and $0.0303 - 0.0169 < 0.0157$, so the estimator's standard error is larger than the difference between SPY's realized Sharpe Ratio and the average. In light of this, SPY not being significantly better is absolutely no surprise.

6. Conclusion

The main aim of this thesis was to tackle the issue of excessive parameter fitting of trading strategies. This phenomenon arises from the possibility of being able to backtest millions of trading strategies quickly, as well as a confusion between realized and true Sharpe Ratios. Eventually, it leads to creating strategies with unrealistically high historical Sharpe Ratios that are inconsistent with future performance.

One way to avoid this is to use the hypothesis test I described in Section 4. Given many strategies, testing whether the historical performance of the best one is significantly better than the average performance can help determine whether parameter fitting is beneficial or harmful. The test is designed to work in realistic market conditions, meaning with returns that are not necessarily normal nor serially uncorrelated.

In the case of 18 international stock indices, I applied this test of two different time horizons. Between 1996 and 2003, the Spanish iShare (symbol: EWD) had the best performance, with a realized Sharpe Ratio of 0.498, while between 2004 and 2020, it was the SPY, the stock index of the US with a realized Sharpe Ratio of 0.479. In both cases, the test found that the outperformance of either index is not statistically significant.

References

[1] David H Bailey, Jonathan Borwein, Marcos Lopez de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance, forthcoming*, 2016.

[2] David H Bailey and Marcos López de Prado. The deflated sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014.

[3] David H Bailey and Marcos Lopez de Prado. The sharpe ratio efficient frontier. *Journal of Risk*, 15(2):13, 2012.

[4] Zdravko I Botev, Michel Mandjes, and Ad Ridder. Tail distribution of the maximum of correlated gaussian random variables. In *2015 Winter Simulation Conference (WSC)*, pages 633–642. IEEE, 2015.

[5] John PA Ioannidis. Why most published research findings are false. *PLos med*, 2(8):e124, 2005.

[6] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1):65–91, 1993.

[7] J Dave Jobson and Bob M Korkie. Performance hypothesis testing with the sharpe and treynor measures. *The Journal of Finance*, 36(4):889–908, 1981.

[8] Oliver Ledoit and Michael Wolf. Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859, 2008.

[9] Pui-Lam Leung and Wing-Keung Wong. On testing the equality of the multiple sharpe ratios, with application on the evaluation of ishares. *Available at SSRN 907270*, 2006.

[10] Andrew W Lo. The statistics of sharpe ratios. *Financial analysts journal*, 58(4):36–52, 2002.

[11] Christoph Memmel. Performance hypothesis testing with the sharpe ratio. *Finance Letters*, 1(1), 2003.

[12] Elmar Mertens. Comments on variance of the iid estimator in lo (2002). Technical report, Technical report, Working Paper University of Basel ï¿œ, 2002.

[13] William F Sharpe. Mutual fund performance. *The Journal of business*, 39(1):119–138, 1966.

[14] John Wright, SC Yam, and Siu Pang Yung. A test for the equality of multiple sharpe ratios. *Journal of Risk*, 16(4), 2014.