Does optimized portfolios outperform the naive diversification: Implications from joint tests

Can optimized trading strategies significantly outperform the naive diversification when correcting for data-snooping

SINDRE GUIN TJETLAND
ANDREAS HÅLAND WEHUS

SUPERVISOR
Valeriy Ivanovich Zakamulin

University of Agder, [2020]
Faculty of Business and Law
Department of Economics and Finance

Acknowledgements

This thesis is the final work of our five-year master education taken at the School of Business and Law at the University of Agder. We have chosen a specialisation within financial economics, allowing us to explore econometric statistics as well as deepen our knowledge and understanding of financial theory. The writing of this thesis would not have been possible without the help of our supervisor Valeriy Zakamulin. His teaching within quantitative finance method motivated us to explore the topic of portfolio optimization. We sincerely appreciate his insightful guidance, feedback, and quick response. Additionally, we would like to thank the other finance professors at the University of Agder for their teachings, and lastly, the School of Business and Law, for giving us the necessary skills to complete this thesis.

Abstract

This thesis expands upon the debate surrounding the paper of DeMiguel, Garlappi & Uppal (2009). We investigate the performance of optimized strategies compared to the naive 1/N rule while controlling for data-snooping. Using the Sharpe ratio and the FFC4 alpha as performance measures, we investigate 10 basic portfolio strategies with datasets from the US and Norwegian markets. We attempt to answer two weaknesses of previous studies on the topic by accounting for data-snooping using White's Reality Check (WRC) and the Superior Predictive Ability (SPA) test. In addition, we include the alpha measure in order to account for the established factor premiums present in the datasets. When we conduct joint-tests on the US datasets, most of our findings show little empirical evidence towards at least one active strategy significantly outperforming the naive benchmark. In the joint-test results from the Norwegian datasets, we find evidence towards at least one strategy significantly outperforming the benchmark. Our study highlights the difference in capital markets while also calling into question the value of asset allocation from optimized strategies.

Contents

List of Figures

List of Tables

1. Introduction

Optimization is the method of finding the best possible solution from all feasible alternatives. Within finance, portfolio optimization is the paradigm of investors constructing the optimal portfolio that has the best tradeoff between the reward and risk. Theorists within this topic attempt to solve this by creating strategies that dictate how an investor should allocate his wealth among a given set of alternatives. Strategies are often tested against historical data, using this as a simulation for how it could perform in a present or future scenario. Several academic papers have attempted portfolio optimization with diverging results. Furthermore, the potential weaknesses of these studies have been highlighted in more recent years, necessitating the need for additional research.

Within finance, portfolio optimization is a topic that has experienced changes throughout the years. One of the first wealth allocation methods is dated back over 1500 years ago. This allocation method stated that one should allocate one-third in land, one-third in merchandise, and one-third as cash. This is one of the earliest example of the naive diversification. The investor splits his wealth equally into the number of assets he wishes to invest in. In this paper, this rule is used as a benchmark comparison for all optimized strategies.

Through the years, there has been a considerable amount of scientific contribution, developing additional strategies, each stating how to best invest wealth in the promise of a future benefit. Most notably, Markowitz (1952) developed a method for constructing an optimal mean-variance portfolio. His method emphasises on the mean and variance of expected return, where the expected return is maximized at a given level of risk. This contribution has become a prominent role in the modern investment theory framework, and founded what is considered the modern portfolio theory (MPT). However, complications arose when applying this method to asset allocation. Researchers discovered that wrongly estimating the parameters could consequently cause extreme portfolio reallocation (Beast & Grauer, 1991) and poor out-of-sample performance (Michaud, 1989).

In response to this, subsequent studies have proposed various methods for mitigating these estimation errors and improve the performance of mean-variance based portfolios. Amongst these attributions, we find the min-variance model (Roger Clarke, Harindra de Silva, & Steven Thorley, 2006), the max-diversification (Choueifaty & Coignard, 2008), the equal-weighted risk contribution (Maillard, Roncalli & Teiletche, 2010), risk parity (Quian, 2011), reward-to-risk and volatility timing strategies (Kirby & Ostdiek, 2012).

DeMiguel et al. (2009) conducted a study where they wanted to evaluate the out-of-sample performance of the sample-based mean-variance strategy, and its extensions designed to mitigate the estimation error, relative to the naive benchmark. Out of their 14 models evaluated across 7 empirical datasets, they find that none of their strategies consistently outperformed the naive diversification. The study came to be regarded as highly influential and has been cited in numerous articles, where financial researchers have attempted similar studies in order to disprove the findings and defend the optimization strategies viability. See; Kritzman, Page & Turkington (2010), Tu & Zhou (2011), Kirby & Ostdiek (2012) and Banerjee & Hung (2013).

Even though the authors mentioned above have implemented improvements to the mean-variance (MV) framework and provided evidence of its efficiency, the extent to which their results are free from data-snooping bias remains unclear. Data-snooping bias arises when researchers apply many empirical tests on the same dataset. White (2000) explained that looking long and hard enough at any given dataset will produce favorable results but that these are, in fact, useless. He thereby argues that data-mining bias is present, when researchers evaluate strategies individually, not accounting for other strategies being evaluated simultaneously. He presents a joint test, called the White reality check (WRC), where potential data-mining bias is reduced. Hansen (2005) proposes an improvement to this test, refereed to as the Superior predictive ability (SPA) test. Yang et al. (2018) and Hsu et al. (2018) addresses the data-mining bias by implementing both the WRC and SPA tests. Both papers conclude that by controlling for data-mining, none of their optimized strategies significantly outperform the naive diversification.

In addition to the data-snooping bias not being accounted for by earlier authors, Zakamulin (2017) discusses an additional weakness in the approach of earlier research. Firstly, he refers to Kritzman et al.; and Kierby & Ostediek for not accounting for the low-volatility effect present in the data. This is also previously discussed by Haugen

& Baker (1991) and Blitz & Van Vliet (2007). Secondly, he discusses the disadvantage of using the Sharpe ratio as the only measurement-tool of performance. He gives a cautionary note regarding the established factors premiums in the datasets of Kenneth French. When controlling for the low-volatility effect, the optimized strategies no longer outperformed the naive 1/N. Sharpe ratio does not take these factor premiums into account when measuring performance.

This thesis attempts to further the discussion of portfolio optimization, where we cover two weaknesses present in previous research on this topic: The presence of data-mining bias and the usage of Sharpe ratio as the only performance measure. We attempt to replicate the findings of both Yang et al. (2018) and Hsu et al. (2018), and also extend it to the Norwegian market. Similarly to these papers, we include the WRC and SPA tests in order to mitigate the data-mining bias. Additionally, we include a second performance measure, the Fama-French-Carhart (FFC4) alpha (Carhart, 1997), as an answer to the established factor premiums, like the low-volatility effect, present in close to all datasets provided by the online library of Kenneth French (Zakamulin, 2017).

The following main null hypothesis is tested: The performance of the best optimized strategy is not significantly better than that of the naive benchmark. If the WRC or SPA tests provides a significant p-value, this signifies the rejection of the null. In this case we see evidence towards the alternate hypothesis, where at least one optimized strategy significantly outperforms the benchmark. A 5% significance level is employed. In this study we apply 10 different portfolio optimization strategies to 16 different datasets based on the US market, and 4 datasets on the Norwegian market. The US datasets contains the period 1963 to 2019, and the Norwegian is for 1980 to 2018.

Our results mostly support that no optimized strategy outperforms the naive 1/N in the US market, with some difference due to the different methodologies in the computations of the performance measures and joint test p-value. For the Norwegian market, we find the opposite results, with the majority of tests pointing towards one or more strategy outperforming the naive.

The remainder of the thesis is structured as follows: Chapter 2 presents a literature review of existing theory. Chapter 3 addresses the methodology applied in the empirical analysis. In Chapter 4, we present the data employed in the study. Chapters 5-6, covers the presentation of the results and followingly, a discussion of these. Lastly, Chapter 7 concludes the thesis.

2. Literature review

The introduction of the mean-variance theory by Markowitz (1952), also known as the modern portfolio theory (MPT), is considered to be a keystone theory when looking at finance and investment. The approach of portfolio optimization suggested by Markowitz relies on estimation of the mean and covariance-variance matrix of a portfolio. With these variables, it seeks to maximize the expected returns at a given level of risk.

Even though the mean-variance (MV) approach suggested by Markowitz has gained much positive recognition, it is not without criticism. Michaud (1989) points out some of the negative sides by applying the MV approach: (i) One of the fundamental problems being that the level of mathematical sophistication required for the optimization is far greater than the level of information in the input forecast. (ii) MV optimizers operate in such a manner that they magnify the errors with the input estimation, Michaud himself referring to the MV model as "error maximizers" due to the large errors following the estimation of mean return and the covariance matrix. Where estimation will always be a problematic aspect of portfolio optimization. He cautions readers of the results from the MV model, as it treats the input variables as true, and not as estimators which they are. Motivated by the inability to accurately predict mean returns, several authors proposed alternative strategies that could reduce the impact of estimation error, briefly mentioned below.

Clarke, De Silva, & Thorley (2006) proposed the minimum-variance portfolio. They focused on risk diversification when developing the strategy, the purpose of the portfolio optimization was to minimize the variance of returns. They applied this strategy to US equity market data provided from the CRSP database, using a time period from 1968-2005. Comparing the performance of the minimum-variance to the market, they provide evidence that the minimum-variance produced higher mean returns, and lower standard deviation.

Seeking to maximize diversification of risk, Maillard, Roncalli, & Teiletche (2009) created a portfolio optimization approach that seeks to equalize the risk contribution (ERC) from each component. Even though the ERC approach was not new, and had already been discussed (Neurich [2008], and Qian [2005]), Millard et al. added new features which focus on single and join risk contribution of the asset. By creating 4 risky assets with 10%, 20%, 30%, and 40%, they applied the naive diversification, minimum variance and the ERC strategies. The result showed that the ERC has lower volatility than the naive diversification, but higher volatility than the minimum variance, but demonstrate that the risk contribution from the minimum variance is less diversified then the ERC. Making the ERC more balanced in terms of risk contribution. They also provide evidence that the ERC produces the highest Sharpe ratio return.

Chaves, Hsu, Li & Sgakernia (2012) discussed the usage of risk-parity for equalizing the risk allocation amongst all assets. Using the classical 60/40 market portfolio as an example, they explain that the 40% containing stocks provide a significant amount more volatility than the 60% containing bonds. They employed datasets provided from S&P 500 from 1980 - 2010. They compared the performance of the risk parity to the 60/40 allocation, the US pension model portfolio (60/40 with anchor), equally-weighted portfolio-, minimum variance- and mean-variance optimization. Their result showed that the risk parity model produces a higher Sharpe ratio than the minimum variance and mean-variance optimization, but fails to consistently outperform the equally weighted portfolio and 60/40 equity bond portfolio. In addition, it has less volatile performance characteristics. They noted that the risk parity is sensitive to the inclusion decision for assets, where including more asset alternatives does not lead to better performance.

Choueifaty & Coignard (2008) wanted to investigate the properties of diversification as a criterion for constructing a portfolio. They created the most diversified portfolio strategy (also referred to as maximum diversification portfolio), which seeks to maximize the diversification ratio of a portfolio — using the S&P 500 index and Dow Jones Euro Stoxx Large Cap index data, covering 1990-2008 for the US and European markets, respectively. They applied the minimum-variance portfolio, equal-weight portfolio, most diversified portfolio towards a market-cap-weighted benchmark. Their results showed that the most diversified portfolios produce a higher Sharpe ratio than the market cap, lower volatility, and higher return in the long run.

DeMiguel et al. (2009) assessed the out-of-sample performance of the sample-based mean-variance model, and its extensions, relative to the naive diversification. Doing so, they applied 14 different portfolio optimization strategies on 8 different datasets provided by Kenneth French. The datasets contained monthly excess returns over the 90-days nominal US T-bill. To measure the performance of the active portfolios, they employed 3 performance measurements: (i) the out-of-sample Sharpe ratio; (ii) the CEQ (certainty-equivalent) return for the expected utility of a mean-variance investor; and (iii) the turnover (trading volume) for each portfolio strategy. The first contribution from their paper is that according to all three of their performance measures, none of the optimized strategies consistently outperform the naive 1/N. Secondly, based on their parameters calibrated to the US market, they showed that the estimated window needed for the sample-based mean-variance strategy and its extensions to outperform the naive diversification is around 3000 for a portfolio with 25 assets, and 6000 for a portfolio. Their result started a "heated debate" on the subject of portfolio optimization, where newer academic papers have attempted to test these findings, employing new allocation methods and datasets, often in order to advocate for a specific strategies viability.

One of the earliest responses that sought to defend the value of optimization is from Krietzman et al. (2009). They pointed out that the equally weighted diversification strategy assumes the investor has no knowledge of the assets. Claiming that if there is some information about the expected return, riskiness, and diversification properties, then the performance gap between the optimized and the naive 1/N would decrease, and the performance of the optimized increase. By using 13 datasets, they constructed more than 50,000 optimized portfolios and evaluated their out-of-sample performance. They showed that the optimized portfolios significantly outperform the naive 1/N, saying: "We found that even without any ability to forecast return, optimization of the covariance matrix by itself adds value." Krietzman et al. (2009) does not check if there is a significant difference between the Sharpe ratios of the strategies, making their statement of outperformance more of an arbitrage observation rather than empirical evidence.

Tu & Zhou (2011) developed an alternative approach to beat the naive 1/N. They proposed a portfolio strategy composed of the naive benchmark together with one of four optimized strategies, where the weights are combined. They found that these combinations have a significant impact on the active strategies, improving the effectiveness compared to their none-combined counterparts. Also, they showed that these combined

strategies significantly outperform the naive 1/N, using the Sharpe ratio as a performance measure.

Kirby & Ostdieks (2012) criticized DeMiguel et al. (2009) and suggested that the high performance of the naive 1/N compared to the active portfolios in the out-of-sample test, is due to their research design. They stated that the focus on portfolios that are subject to high estimation risk and extreme turnover favors the naive 1/N strategy. When applying their own model, they discovered that the mean-variance optimization often outperforms the naive 1/N but see that turnover can erode its advantage in the presence of transaction cost. Turnover is defined as how quickly company receives cash from accounts receivable, or how quick they can sell its inventory. Solving this problem, they developed two new methods of mean-variance portfolios, characterized by a low turnover. The two new strategies are called volatility timing and reward-to-risk timing, based on the earlier work of Fleming, Kirby & Ostdieks (2001, 2002). Some key features of these strategies which give them an advantage are like those we can find in the naive 1/N. (i) No optimization, (ii) no covariance matrix inversion, and (iii) no short sales. Implementing the two strategies shows that they outperform the naive 1/N portfolio.

Banerjee & Hung (2013) compared the active momentum trading versus the naive 1/N strategy, aiming to discover the merits and demerits of the momentum trading. By looking at the time period from 1926 to 2005 and 100 randomly selected 10-year periods, they implemented the active momentum trading versus the naive 1/N strategy. When comparing the difference between the profits of momentum and the naive benchmark, they find that there are no significant differences.

More recently, Zakamulin (2017) wrote a cautionary note regarding the usage of Kenneth French's datasets while applying the Sharpe ratio as a performance measure. After examining earlier work related to the response of DeMiguel et al. (2009), Zakamulin pointed out similarities amongst Kritzman et al. (2010), Tu and Zhou (2011) and, Kirby and Ostdiek (2012). (i) They all implemented various portfolio optimization methods using datasets provided by Kenneth French, and (ii) the usage of the Sharpe ratio as a performance measurement without controlling for the risk-based explanation of the superior performance of their optimized portfolios.

Zakamulin points out that the authors do not consider the possibility that some profit anomalies may provide the superior performance of the optimized strategies seen in some papers. Testing the 17 different datasets provided by Kenneth French, he proves

the existence of low-volatility effects in virtually all of them. The alphas generated by the CAPM-model returns statistically significant values in his findings. However, by expanding upon the alpha, and adding the Fama-French HML factor to the model as a way to control for the low volatility effect (Blitz, 2016), there appears no evidence of the optimized strategies outperforming the naive. Therefore in order to accurately measure and test for outperformance, the alpha measure should be used, either from the Fama-French 3-factor model (FF3) or Fama-French-Carhart 4-factor model (FFC4).

Another criticism to Kritzman et al. (2010), Tu and Zhou (2011) and, Kirby and Ostdiek (2012), is that these articles do not control for data-mining in their datasets. Data-mining occurs when re-using the same dataset multiple times. When re-using the dataset, there is a possibility that the satisfactory result is more chance, rather than from a superior underlying methodology of the allocation strategy. White (2000) proposed a method for countering the effect of data-mining bias, White's Reality Check. With the purpose of creating a method for testing the null hypothesis that the best model encountered has no predictive data-mining bias superiority over a given benchmark. This joint test allows with some degree of confidence, the conclusion that the results are not from pure chance or luck, but rather true outperformance. The WRC test incorporates all optimization models and datasets into one joint test, for a final p-value. Hansen (2005) extended this model and proposed minor changes to increase its overall adequacy; his extension is referred to as the SPA test for Superior Predictive Ability. It differs from the WRC in that it normalizes the test statistic, and use a threshold for the z-values, removes inferior performing strategies from the model altogether.

The article of Hsu, Han, Wu, & Cao (2018) in addition to the one by Yang, Cao, Han & Wang (2018) both employed the joint SPA and WRC test to correct for the data-mining issue present in earlier studies. Thus more accurately determining the performance of some active trading strategies compared to a naive 1/N benchmark. Hsu et al. (2018) applied 16 different strategies to 3 different datasets, finding limited evidence that the active strategies outperform the naive. Cao et al. (2018) applied 11 different portfolio strategies to 4 different datasets, comparing the strategies to the naive, where they similarly find little evidence of their tested strategies outperforming the naive. They expand upon the joint tests by implementing step-wise iterations. Thereby they can determine how many strategies potentially outperforms the naive. Both articles highlight the significant impact data-mining bias has on the overall conclusion of a study.

3. Methodology

3.1 Active strategies

This section provides a description of the methodology used in the empirical analysis of the study. Presenting a framework for the strategies, joint tests, backtesting and performance measures. In Table 3.1 we list the strategies used, with their abbreviations.

Table 3.1: Allocation strategies

#	Model	Abbreviation
	Benchmark	
0.	1/N	NAIVE
	Active strategies	
1.	Mean-variance	MV
2.	Minimum variance	MVP
3.	Volatility timing	VT
4.	Reward to risk timing	RRT
5.	Equal risk contribution	ERC
6.	Risk parity	RP
7.	Minimum tail dependent portfolio	TD
8.	Risk efficient portfolio	RE
9.	Maximum diversification	MDIV
10.	Maximum-decorrelation	MDEC

Table 3.1 lists the various allocation strategies used. # denotes the strategy number, and the last column the strategy abbreviation.

This paper uses 10 portfolio optimization strategies in addition to the naive benchmark strategy. The strategies employ a combination of historical returns, variance-covariance matrix, and the risk-free rate of return. These values are used in order to compute the allocated weights for each asset i. The strategies, therefore, present an optimization problem with regards to the weights, limited by two constraints. The first constraint is that the total sum of weights equals to 1, and the second constrains is that short sales are not allowed. Where 3.1 represents the sum of all the weights, and 3.2 the short sale restriction.

$$\mathbf{w}'\mathbf{1} = 1 \tag{3.1}$$

$$w_i \geq 0, \tag{3.2}$$

where w_i is the wight of assets i in regards to the total portfolio weights $\mathbf{w}$, $\mathbf{w} = (w_1, w_2, ..., w_N)$

Throughout this section, we employ several symbols that are used repeatably for all strategies. μ denotes the expected mean return of a portfolio, and σ is the standard deviation. $\mathbf{w}'\mathbf{\Sigma}\mathbf{w}$ is the variance of a portfolio, and $\mathbf{\Sigma}$ is the $N \times N$ variance-covariance matrix of returns. Matrices and vectors are bolded.

(0) 1/N benchmark

The naive diversification is a well-known investment strategy that has been used for many years, often as a benchmark comparison (DeMiguel et al. 2009). Naive diversification, or sometimes referred to as 1/N rule, consists of dividing the wealth equally amongst the N risky assets. It does not consider historical returns or risk, separating itself from the other strategies by being the simplest strategy and having no optimization. In this paper, the 1/N strategy is used as a benchmark for comparison against active strategies.

$$w_i = \frac{1}{N} \tag{3.3}$$

(1) Mean-variance

The mean-variance strategy analyses the risk against the expected return. This strategy was proposed by Markowitz (1952). The strategy helps the investor find the portfolio which give the highest return at a given level of risk. This is done by using the estimated mean return and the variance-covariance matrix of the returns.

The strategy relies on investors behaving rationally, that an investor choosing between two portfolios with similar returns, but different variance, will choose the portfolio with the lowest risk. The asset weights of the strategy is found by maximizing the expression below:

$$\max_{w} \mu - \lambda \mathbf{w}'\mathbf{\Sigma}\mathbf{w} \qquad s.t. \qquad \mathbf{w}'\mathbf{1} = 1\ , \quad w_i \geq 0, \tag{3.4}$$

where λ is the parameter which measures how risk averse an investor is, it is set at: 0.89. There is no general solution to this problem, and it is solved with the use of a quadratic solver from the package "optimalPortfolio"(2020) in R.

(2) Minimum variance portfolio

The mean-variance portfolio, amongst other "optimized" portfolios, is constructed upon estimated risk and return. Estimated returns will inevitably become subject to estimation error to a higher degree than estimated risk (Michaud, 1989). The minimum variance only depends on the covariance matrix and is, therefore, less exposed to estimation errors. The MVP is located on the left-most area of the efficient frontier, where risk in the form of variance is minimized. It was shown by Clarke, DeSilva, and Thorley (2006) that the MVP strategy achieves both low risk and a high Sharpe ratio marking the strategy as attractive for investors.

In order to estimate the MVP portfolio, one needs to calculate the weights of each asset i. In this paper, the weights are calculated with short sales restrictions. The following minimization problem is presented in order to compute the weights:

$$\min_{w} \mathbf{w}'\mathbf{\Sigma}\mathbf{w} \qquad s.t. \qquad \mathbf{w}'\mathbf{1} = 1\ , \quad w_i \geq 0 \tag{3.5}$$

The problem is solved by using the "optimalPortfolio" package in R.

(3) Volatility timing

Kirkby and Ostdiek (2012) proposed a set of timing strategies where the investors consider mean returns $\boldsymbol{\mu}$ constant and equal: $\mu_i = \mu_j = \boldsymbol{\mu}$. The weights are therefore determined by the conditional covariance matrix, not the expected return. They find that the volatility timing strategy (VT) overall outperforms mean-variance efficient portfolios. The weights of asset i are given:

$$w_i = \frac{\left(\frac{1}{\sigma_i^2}\right)^{\eta}}{\sum_{i=1}^{N}\left(\frac{1}{\sigma_i^2}\right)^{\eta}}, \tag{3.6}$$

where σ_i^2 is the $i-th$ standard deviation asset i. $\eta > 0$ is a tuning parameter, measuring the timing aggressiveness of the strategy; in other words, how sensitive the weights are to volatility changes.

In this paper $\eta = 4$, this in accordance with Zakamulin (2017) and the findings of Kirby & Ostdiek (2012). The latter, who discovered that setting $\eta > 1$ should help compensate for the information loss caused by ignoring the correlations.

According to Kirby & Ostdiek (2012), 4 notable features characterize the VT strategy: (i) First, it does not require optimization, (ii) Second; it does not require covariance matrix inversion, (iii) Third, they do not generate negative weights, (iv) Fourth, through volatility changes, the sensitivity of the portfolio weights can be adjusted with a tuning parameter.

(4) Reward to risk timing

Kirby & Ostdiek (2012) also propose a more general timing strategy that incorporates the expected return μ and thus attempting to better capture all factors in the optimization. Similarly to VT, reward-to-risk timing (RRT) ensures non-negative weights of asset i. The weights are given by:

$$w_i = \frac{\left(\frac{\mu_i^+}{\sigma_i^2}\right)^{\eta}}{\sum_{i=1}^{N}\left(\frac{\mu_i^+}{\sigma_i^2}\right)^{\eta}}, \tag{3.7}$$

where μ_i^+ is the estimated mean of the returns subtracted the risk-free rate, with

the limitation that the result is positive, assuring non-negative weights. The tuning parameter is like the VT strategy set at 4 in this paper. σ_i^2 is the variance of asset returns. The reward to risk strategy includes returns in its calculations. These are often estimated with less precision than the variance. In order to reduce this effect, the mean excess returns in the formula above are limited in only positive returns; $\mu_i^+ = \max(\mu_i, 0)$. The strategy is computed in R.

(5) Equally risk contribution portfolios

Equal Risk Contribution (ERC) is a risk-balanced portfolio proposed by Maillard, Roncalli & Teiletche (2008). When an asset produces significantly higher risk than another, the weight of the high-risk asset will get scaled down so that the relative risk contribution will be on an even level compared to the less risky. This requires the estimation of the covariance matrix for computation of the weights. The relative risk contribution (RRC) is normalized and shows the risk contribution of asset i in the portfolio:

$$RRC_i = \frac{w_i(\mathbf{\Sigma w})_i}{\mathbf{w'\Sigma w}}, \tag{3.8}$$

where $(\mathbf{\Sigma w})_i$ denotes the element i of vector $(\mathbf{\Sigma w})$ Satisfying that the sum of all RRC is equal to 1.

The overall goal of the ERC is to equalize the relative risk contribution amongst N assets such that:

$$RRC_i = \frac{1}{N} \tag{3.9}$$

The ERC portfolio equalizes RRC amongst all assets and presents itself as an optimization problem:

$$\min_{w} \sum_{i=1}^{N}\sum_{i=1}^{N} \Big(w_i(\mathbf{\Sigma w})_i - w_j(\mathbf{\Sigma w})_j\Big)^2 \qquad s.t. \quad \mathbf{w'1} = 1\ , \ \ w_i \geq 0 \tag{3.10}$$

It is solved using an SQP (sequential quadratic programming) algorithm, from the R package FRAPO (2016).

(6) Risk parity

The risk parity model is a simplified version of the ERC model, proposed by Asness, Frazzini, & Pedersen (2012). This strategy does not use the variance-covariance matrix when calculating weights. Instead, it relies on the standard deviation of the returns as a measure of risk. Assuming zero correlation between the assets $\rho_{ij} = 0$, the weights are calculated as shown:

$$w_i = \frac{\frac{1}{\sigma_i}}{\sum_{i=1}^{N} \frac{1}{\sigma_i}} \tag{3.11}$$

Due to the equal distribution of risk, less risky assets are therefore overrepresented relative to the market portfolio, reducing the overall risk of the risk parity portfolio. The strategy is computed in R

(7) Minimum tail dependent portfolio

The Minimum Tail Dependent (MTD) portfolio is a non-parametric portfolio estimation strategy where the variance-covariance matrix is replaced by a matrix of the lower tail dependence coefficients (Adame-Garcia, Fernandez Rodriguez, & Sosvilla-Rivero, 2017). Instead of minimizing volatility as in the MVP portfolio, the MTD portfolio attempts to minimize tail dependence. Tail dependence is an expression of the dependence of the relationship between extreme returns of N assets. In other words, it is a measure of external dependence on the returns. With lower tail dependence, we are considering extreme negative returns (Schmidt, Stadtmuller, 2006). The overall weights optimization is akin to the method used in the global minimum variance portfolio, differing in the matrix used.

$$\min_{w} \mathbf{w'TDw} \qquad s.t. \qquad \mathbf{w'1} = 1\ , \quad w_i \geq 0, \tag{3.12}$$

where **TD** is the tail dependence matrix of the returns. This minimization problem is solved using the R package FRAPO (2016) FRAPO uses a non-parametric estimation provided by Schmidt et al. (2006) in order to compute the lower tail dependence. (Nelsen, Quesada-Molina, Rodriguez-Lallena & Ubeda-Flores, 2000).

(8) Risk efficient portfolio

Developed by Amenc, Goltz, Martellini, & Retkowsky, (2011), they propose to construct a maximum Sharpe ratio portfolio with the assumption that a stock's expected return is a deterministic function of its semi-deviation, and the cross-sectional distribution of semi-deviations. The following maximization problem is presented:

$$\max_{w} \frac{\mathbf{w'J\xi}}{\sqrt{\mathbf{w'\Sigma w}}} \qquad s.t. \qquad \mathbf{w'1} = 1\ , \quad w_i \geq 0, \tag{3.13}$$

where $\mathbf{J}$ is a $(N \times 10)$ matrix of zeroes who's (i,j)th element is 1 when the semi-deviation of asset i belongs to decile j. For each decile j; $\xi \equiv (\xi_1, ..\xi_{10})$' the mean semi-deviation is then calculated and used as the expected return. Instead of using expected returns in the optimization, this strategy creates decile portfolios with regards to the stocks semi-deviation.

(9) Maximum Diversification

Choueifaty and Coignard (2008) suggested the most diversified portfolio strategy. It seeks to maximize a portfolio's diversification. The strategy implements the diversification ratio proposed by Choueifaty (2006), defined as the ratio of the portfolio's weighted average volatility to its overall volatility.

$$DR\,(w) = \frac{\mathbf{w'\boldsymbol{\sigma}}}{\sqrt{\mathbf{w'\Sigma w}}} \tag{3.14}$$

We maximize the ratio with regards to $\mathbf{w}$ in order to compute the portfolio weights:

$$\max_{w} \frac{\mathbf{w'\boldsymbol{\sigma}}}{\sqrt{\mathbf{w'\Sigma w}}} \qquad s.t. \qquad \mathbf{w'1} = 1\ , \quad w \geq 0, \tag{3.15}$$

where $\boldsymbol{\sigma}$ is a vector of asset volatilities: $(\sigma_1 ... \sigma_N)$. A DR ratio of a long-only portfolio is greater or equal to one, due to *" the volatility of a long-only portfolio of assets is less than or equal to the weighted sum of the assets' volatilities"* (Choueifaty, Froidure, Reynier, 2011). The DR ratio is maximized with regards to the weights and computed with R package "RiskPortfolios" (2020). '

(10) Maximum-decorrelation

Christoffersen et al. (2012) presented an approach where the portfolio volatility is minimized under the assumption that individual asset i variance is identical. This approach only relies on the asset's correlations, attempting to exploit the connection between correlation and the asset volatility in order to minimize overall risk through decorrelation. The maximum-decorrelation strategy is a direct application of this approach, and suggested by Goltz & Sivasubramanian (2018).

Therefore, assuming equal volatilities across all assets; $\sigma_i = \sigma_j = \sigma$ Where $\boldsymbol{\sigma}$ is assumed to be constant. The $\boldsymbol{\Sigma}$ covariance matrix is replaces with the correlation matrix of the assets, $\mathbb{R}$ in the minimization problem. The expression for optimal weights is then derived, minimizing the expression:

$$\min_{w} \mathbf{w}'\mathbb{R}\mathbf{w} \qquad s.t. \qquad \mathbf{w}'\mathbf{1} = 1\ , \quad w \geq 0, \tag{3.16}$$

where $\mathbb{R}$ is the correlation matrix of the assets. The weights form this optimization problem is estimated with the "optimalPortfolio"(2020) package in R.

3.2 Backtesting

Backtesting, also called out-of-sample simulation, is the process of applying trading strategies to historical data and also measuring their performance. Doing so it simulates how the strategies would have performed in "real-life". An investor has at the current time t information about historical data up to present time t, but not the future $t+1$ and beyond. Using the data available from time period 1 to t, one can estimate the weights and followingly the returns of the various allocation strategies employed in the simulation. At the current time t historical data before time t is employed to predict future data $t+1, t+2 ... t+N$ This corresponds to respectively "in-sample" and "out-of-sample" time windows. The "out-of-sample" estimation becomes the future forecast of a portfolio manager's weight optimization.

The rolling window estimation method uses the "in-sample" and "out-of-sample" windows to calculate parameters. For each new time period added to the estimation, $t+1, t+2 ... t+N$, the rolling window estimation removes the earliest period. Thus

a rolling window refers to the movement of the "in-sample". For the whole dataset, a continually shifting "in-sample" window calculates "out-of-sample" results, weights, returns and the variance-covariance matrix. This process is continued until we have estimated out-of-sample returns for the $T - M$ period. Where M is the lookback, period and T is the total timespan.

In this thesis, a rolling window of 60 months is used (lookback period). This value is chosen to correspond with the findings of DeMiguel et al. (2009), where they find that a time window of 60 months does not produce largely different results than a time window of 120 months.

3.3 Performance measures

In order to measure the performance of our chosen portfolio strategies, we consider two distinct measures: The FFC4 alpha model, and the Sharpe ratio. For each individual strategy, we measure the performance corresponding to the number of datasets. Thereafter, the performance measures of the active strategies are compared to those of the naive 1/N. The Sharpe ratio is included in this thesis on the basis of DeMiguel et al's. (2009) paper, allowing easy comparison of results. In addition, the alpha measure is included based on Zakamulin's (2017) response to Kritzman et al. (2010) and Kirby & Ostdiek (2012), where he points out the weakness of having the Sharpe ratio as the single performance measure.

Sharpe ratio

The Sharpe ratio was introduced by Sharpe (1996) and is a risk-adjusted performance measurement often used by financial researchers to evaluate portfolio strategies. It compares the performance of an investment compared to a risk-free asset, after adjusting for the portfolio risk. As a performance measure, the Sharpe ratio is a quick and easy method of comparing different risk exposed portfolios. The simplicity and ability to compare performance makes it a favorable technique. The formula is given by:

$$Sharpe \;\; Ratio = \frac{\mu_p - r_f}{\sigma_p}, \tag{3.17}$$

where μ_p is the return of portfolio strategy p, r_f is the risk-free rate and σ_p the total risk.

While being a widely employed measure, the Sharpe ratio has some argued drawbacks. Firstly, the Sharpe ratio is accentuated by portfolios that do not have a normal distribution. This is due to the model not differentiating between downside and upside risk (volatility), leading to the ratio presenting a misleading image of the performance, where abnormal distribution skews the ratio as more favourable than in reality. An example of this being hedge funds. Secondly, it does not consider the risk-based explanations of the performance.

The Sharpe ratios are compared to each other, testing if the optimized strategies ratios SR_j are statistically distinguishable from the naive benchmarks ratios $SR_{\frac{1}{N}}$. It is tested statistically under the following null hypothesis:

$$H_0 \;:\; SR_j \leq \; SR_{\frac{1}{N}} \tag{3.18}$$

With the test statistic:

$$z \;=\; \frac{SR_j - SR_{\frac{1}{N}}}{\sqrt{\frac{1}{T}\left[2\left(1-\rho\right) + \frac{1}{2}\left(SR_j^2 + SR_{\frac{1}{N}}^2 - 2SR_j SR_{\frac{1}{N}}\rho^2\right)\right]}}, \tag{3.19}$$

where ρ is the correlation coefficient, and T is the sample size. Estimating the p-values, we employ a 5% significance level. Rejecting the null if the p-value is below the α.

Alpha FFC4

Zakamulin (2017) points out the disadvantage of using the Sharpe ratio as the only measurement-tool of performance. The main reason is that a high measured Sharpe ratio may be the result of some known return anomalies. Leading to a performance measure that does not accurately describe the overall performance. His paper provided evidence that the low-volatility effect is present in almost every dataset provided by the online data library of Kenneth French.

Another performance measure to complement the Sharpe ratio is the alpha. It is an independent variable that captures the effect of some known return-anomalies. In order to capture as much as possible of the abnormal market return, the Four-Factor-Carhart (FFC4) model is applied in the thesis.

$$R_{pt} = \alpha_p + \beta_1 Mkt - r_f + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 MOM_t + \epsilon_t, \tag{3.20}$$

where, $E[\epsilon] = 0$.

This model is a further extension of the Fama & French FF3 model (Carhart, M, 1997), and includes the additional factors added by this model compared to the simpler CAPM, or the FF3 model (Fama & French, 1993). In addition, it includes a momentum factor, where the original alpha is expanded to now include four factors to better capture the abnormal performance compared to a benchmark market. Explained, R_{pt} is the expected return on the portfolio strategy p. α is the intercept capturing the abnormal performance of the tested strategies. The first of the 4 factors are $Mkt - r_f$, which is the excess market risk premium. SMB which is the expected return of the value factors. The SMB factor stands for Small Minus Big (market capitalization), and reflect the average returns from the three smallest portfolios minus the average returns from the three largest. HML is the expected return on the size factors. The HML factor stands for High Minus Low and accounts for the spread in returns between value stocks and growth stocks. Lastly MOM is the expected return on the momentum factors, it is described as the tendency of a stock to continue increasing if it is going up, and continue declining if it is going down. This phenomenon can be measured by subtracting the equally-weighted average of the "losing" stocks from the equal-weighted "wining" stocks. The beta coefficients β_1, β_2,β_3 and β_4 denoted the exposure to the respective factors. In an adequate model, these factors will explain the performance accurately such that

the alpha measures only measure the abnormal performance of an allocation strategy. A high alpha signifies the strategy beating the market and is desirable.

We employ the alpha measure in a test, whether it is statistically different from the naive. Calculating the alpha for both the optimized strategies and the naive benchmark and testing for the difference. Formulated as a hypothesis:

$$H_0 : \alpha_j \leq \alpha_{\frac{1}{N}}, \tag{3.21}$$

where α_j denoted the alpha for the optimized strategy j, and $\alpha_{\frac{1}{N}}$ denoted the alpha for the naive 1/N benchmark. If the p-value is below a significance level of 5% we reject the null and conclude in a significant difference between the measure. With the test statistic:

$$z = \frac{\alpha_j - \alpha_{\frac{1}{N}}}{\sqrt{\frac{1}{T}(\sigma_j^2 - 2\rho\sigma_j\sigma_{\frac{1}{N}} + \sigma_{\frac{1}{N}}^2)}}, \tag{3.22}$$

where ρ is the correlation coefficient, and T is the sample size. σ is the standard deviation.

The low-volatility effect present with the Sharpe ratio has been found to disappear when measuring performance with a multi-factor alpha. Specifically, Scherer (2011) concludes that the High Minus Low (HML) factor from Fama & French explains some of the superior performance seen in an optimized portfolio. Zakamulin (2017) further discusses this phenomenon, stating that the HML factor controls for the low volatility-effect presented in datasets.

Both the Sharpe ratio and alpha offers a way to measure expected return on a risk-adjusted basis. An aspect of the alpha is that it a useful measurement tool when measuring performance in relation to the market. If α_i is equal to 1, then the active strategy measured beats the market by 100 %.

3.4 Data-mining bias

Data-mining is the process of finding the best trading strategy among a great number of alternatives. The strategy selected is the one with the best-observed performance in the backtest. An issue with data-mining is that the best strategy rule systematically overestimates the true performance of the strategy. Referred to as a data-mining bias, it is linked to the randomness of returns estimated in the backtest. If we define returns as performance, it can be separated into two components, true performance and randomness:

$$Observed \;\; performance = true \;\; performance + randomness \tag{3.23}$$

The random component in the equation can manifest as either "good luck" or "bad luck". Where the true outperformance benefits from "good luck" and diminishes from "bad luck". Suppose that the true performance of the active trading strategy is equal to that of the benchmark. Then the expression would become:

$$Observed \;\; outperformance = 0 + randomness \tag{3.24}$$

Where it is the randomness factor that now defines the optimized strategies and outperformance compared to the naive, this however, presents the data-mining-issue. If true performance = 0, then outperformance = randomness. It is found that when data-mining for the best strategy in a backtest, the strategy selected is systematically the one with the highest positive randomness — in other words, attributed to luck (Zakamulin, 2017).

When analyzing a single strategy from the backtest, and determining if the measured performance is significant, the p-value is given:

$$p_S = Prob\left(z > z_{1-\alpha}\right), \tag{3.25}$$

where $z_1 - \alpha$ is the $1 - \alpha$ quantile of the standard distribution. In our thesis we have set the significance level at $\alpha = 0.05$, then $p_S = 5\%$. This means that there is

a 5% chance of false a discovery in a single test. The problem arises when analyzing several strategies in several datasets. The 5% chance is applied to each test, but when performed N times, the chance for the type 1 error goes up when looking at all datasets and strategies combined.

If we assume the test statistics of the N strategies are independent, and that for all strategies true outperformance equals to zero; The probability that when testing several strategies, at least one p-value is below the significance level of 5% is given by:

$$p_N = 1 - prob\left(z_j < z_{1-\alpha}; z_{\frac{1}{N}} ...; z_N < z_{1-\alpha}\right) = 1 - (1 - p_S)^N , \tag{3.26}$$

where z_i, is the value of test statistic i, tested against the test statistic $z_{\frac{1}{N}}$, with a significance level $z_1 - \alpha$. Note that the p-value P_N, is one minus the probability that in all the independent test, the p-values are less than α. This is because a single test $p_S = (z_i < z_1 - \alpha) = 1 - Ps$ and all test are independent when finding the probability of N independent test, and all p-values are less than α equals $(1 - p_S)^N$.

The result of this equation is that when testing several strategies as in this thesis, and assuming a true outperformance of 0, the chance for at least one of these being below our significance level is very high. This is then attributed to the randomness factor of the outperformance, meaning that the luckiest strategy is presented as the winner. The chance of finding at least one strategy which "outperforms" the benchmark increases with adding more strategies, N. Using $N = 10$ gives $1 - (1 - 0.05)^{10} = 0.401$. 40.1% chance of at least finding one strategy that "outperforms". If $N = 100$, then $1 - (1 - 0.05)^{100} = 0.994$. This implies that there is a probability close to 100% of finding at least one strategy that "outperforms". In order to combat the data-mining bias we implement a combined significance test on the performance measures instead of several independent ones, reducing the chance of incorrectly rejecting the null.

3.5 Combined test

The two performance measures alpha, and Sharpe needs to be tested to determine if the values are statistically significant. Individual tests investigate the significance of optimized strategies against the naive, performing a null-hypothesis versus an alternative-hypothesis on all the strategies for each of the datasets. This however, gives way to a data-mining issue as discussed in the relevant section above. In this thesis, we additionally employ the use of two statistical testing methods where the combined excess returns from all strategies and datasets are tested. This ensures that data-mining bias is minimized.

Bootstrap

Bootstrapping is a method of recreating datasets. What the bootstrap does is quantifying the uncertainty by estimating confidence intervals, standard errors, and performing significance tests. Compared to more traditional methods of estimation, the bootstrap requires fewer assumptions and, in general, produce results with greater accuracy. These assumptions are, for instance, that the datasets must be normally distributed, and the number of observations must be more than 30. The bootstrap method is a popular tool since it does not require any parametric assumptions on the random variables and can be applied to smaller datasets (less than 30). However, in this thesis, the data size will not be a problem since the US dataset contains over 600 observations and the Norwegian over 432 observations.

The bootstrap-method studies the statistical inference of variables. Statistical inference is based on the sampling distribution of a sample statistic. When conducting statistical inference on an estimator (random variable), a requirement is to know its probability distribution. By knowing the distribution of a random variable, one can further estimate its standard errors and confidence intervals. The procedure for bootstrapping a sample is done by firstly creating k new samples, which we will refer to as resamples. This is done by sampling with replacement. By taking the original sample, we randomly extract a variable and use this for our new sample. Each variable can be extracted multiple times. The process is repeated until our resample is the same size as the original sample, and our first resample is complete. The process is repeated k amount of times, where a larger k provides more accurate results. The statistics for each resample is then calculated; this is called a bootstrap distribution. It gives information

about the center, shape, and spread of the random distribution (Hestenberg, Monaghan, Moore, Clipson & Epstein, 2005).

This method is applied to our calculated t-statistics. By resampling our statistics and creating a bootstrap distribution for both the Alpha and Sharp values, we will, with a certain degree of confidence, know if one active strategy outperforms the naive benchmark. White (2000) recommends using a moderately large number of resamples, approximately 500 or 1000. His recommendation has been considered, and the k value (amount of resamples) in our calculation is set equal to 10 000. Increasing the number of resamples will produce a better estimate of the sampling distribution and a more reliable estimation of the standard error.

WRC - White's Reality Check

White (2000) addresses the possibility that any satisfactory result obtained from data-mining may be a result of chance/luck. Solving this problem, he created a method for testing the null hypothesis that the best model encountered in a specific search, has no predictive superiority over a benchmark model. The alternative hypothesis consequently being that there is at least one strategy with greater predictive ability relative to the benchmark model. With the alternate hypothesis that there is at least one strategy with an excess return greater than the benchmark.

Suppose we want to test the performance of N strategies against a given benchmark. Our purpose is to determine if one of our N strategies could consistently outperform the benchmark, at a given confidence level. Using the Sharpe ratio performance measure as an example, the WRC test tests for outperformance.

Outperformance is defined as the excess performance of a strategy j when subtracting the performance of a benchmark. Our thesis compare 10 active trading strategies to the naive 1/N, denoted f_j.

$$f_j = SR_j - SR_{\frac{1}{N}}, \tag{3.27}$$

where f_j is the measured outperformance. SR_j is the Sharpe ratio of allocation strategy j, and $SR_{\frac{1}{N}}$ the benchmark.

The model/strategy with the highest outperformance measure f_j is furthermore chosen, and this models outperformance is denoted $\overline{f}$.

$$\overline{f} = \max_{j=1,\dots,N} f_j \tag{3.28}$$

We want to apply $\overline{f}$ to determine the following null hypothesis: The best model is not better than the benchmark. The alternate hypothesis being that the best model is better than the benchmark:

$$H_o : \overline{f} \leq 0 \quad VS \quad H_1 : \overline{f} > 0 \tag{3.29}$$

A geometric bootstrap method is used to compute the resampled outperformance measures. This is done because $\overline{f}$ is a random variable, and for statistical inference, we need to know its distribution. Doing so, we can estimate its standard error and confidence intervals, which is necessary for construction of the null and alternative hypothesis. A joint sample is conducted of the excess return to the strategies and the naive benchmark. After the bootstrap, the resampled optimized strategies returns are denoted $r^*_{k,j}$ and the benchmark; $r^*_{k,\frac{1}{N}}$. Subscript k implies the $k-th$ iteration of the bootstrap, which is performed k times. Subscript j denotes the optimized strategy number, in this paper a number between 1-10

For each iteration of the bootstrap, the resampled Sharpe ratio for all strategies is estimated using the new returns. Thereafter we compute the outperformance; this is done by subtracting the benchmark Sharpe ratio from the optimized strategy ratio:

$$f^*{}_{k,j} = \ SR^*{}_{k,j} - SR^*{}_{k,\frac{1}{N}}, \tag{3.30}$$

where $f^*{}_{k,j}$ is the resampled outperformance of the model j of iteration k, and $SR^*_{k,j} - SR^*_{k,\frac{1}{N}}$ is the Sharpe ratio of the active strategy subtracted by the Sharpe ratio of the benchmark strategy.

Thirdly, after computing the resampled outperformance measures for each strategy j, we chose the best-observed outperformance for each iteration of k:

$$\overline{f^*_k} = \underset{j=1,\dots,N}{max} \left(f^*_{k,j} - f_j \right), \tag{3.31}$$

where $\overline{f}_k^*$ is the best outperformance measured, and the bootstrapped test statistic. In order to comply with the null hypothesis that no optimized strategy outperforms the naive, the computed outperformance f_j^* is adjusted by subtracting the originally observed outperformance. Thereafter the maximum value is chosen. The collection of $\overline{f}_k^*$ gives the distribution of $\overline{f}$.

The next step is to test whether the chosen maximum outperformance is significantly different from 0. The p-value of the test is computed by counting the number of times $\overline{f}_k^*$ is greater than $\overline{f}$, within the k iterations of the bootstrap.

$$P_{WRC} = \sum_{k=1}^{k} \frac{^{I}\left(\overline{f}_k^* > \overline{f}\right)}{k}, \tag{3.32}$$

where P_{WRC} is the observed p-value of the test, I denotes the indicator function that takes the value of one if the condition is satisfied.

Finally, this gives enough information to check the null hypothesis:

$$H_0 : \overline{f} \leq 0 \tag{3.33}$$

Thereby testing whether outperformance of the optimized strategies used in this thesis outperforms the naive by a significant amount. The significance level is set at 5 %.

SPA - Hansen's test for superior predictive ability

Hansen (2005) suggested two improvements to White's reality check. The first suggestion is that the test statistic should be normalized/studentized. This changes the value of the outperformance f_j, by scaling it down by its standard deviation $\sigma_{f,k}$. The test statistic is given by:

$$z_k \ \rightarrow \ \frac{f_k}{\sigma_{f,k}} \tag{3.34}$$

The second suggestion is the removal of poor performing strategies. Hansen suggests that a threshold value should be introduced, excluding the values below. This value is calculating by the equation:

$$A \;=\; -\sqrt{2\ln\left(\ln\left(n\right)\right)}\,, \tag{3.35}$$

where A denoted the threshold value, and n is the number of observations in the datasets. The following threshold values for the US and Norwegian market are presented in Table 3.2 below.

Table 3.2: Threshold values

Datasets	Observations	Threshold (A)
US	618	-1.928
Norwegian	366	-1.884

Table 3.2 reports the threshold values sorted by the US and Norwegian datasets.

The datasets measured are divided between US and Norwegian datasets. We list the number of observations we have for each dataset, and the threshold value (A) is the cutoff calculated in Equation 3.35 for bad strategies. By comparing the z-stats calculated for each dataset ($\overline{f}$) to the threshold value, we remove all strategies on the US dataset which are below $z_k < -1.928$, and for the Norwegian $z_k < -1.884$.

4. Data

The data employed to investigate the performance of the selected optimized strategies comes from two sources. For the US data, we downloaded our datasets from the online data library of Kenneth French[1]. He supplies a wide range of dataset which are formed on different criteria and are frequently updated. They employ data from NYSE, AMEX, and NASDAQ stock indexes. These datasets are similar to those used by DeMiguel et al. (2009), Kritzman et al. (2010), and Zakamulin (2017). The Norwegian datasets are downloaded from the online data library of Ødegaard[2]. His datasets are created from the Oslo stock exchange data service. The primary data downloaded from these sources are value-weighted portfolio returns formed on different criteria, presented monthly. Several datasets are employed in order to improve the validity of our findings, attempting to emulate real trading markets. Both the US and Norwegian data is used for the purpose of investigating potential similarities and differences in the trading markets.

In this chapter, we present the datasets employed for both markets, 16 for the US and 4 for Norway. The significant difference in the number of datasets analyzed comes from the reduced availability of developed datasets in Norway.

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[2] http://finance.bi.no/ bernt/financial$_data/ose_asset_pricing_data/index.html$

The US data uses inputs from stocks on the NYSE, AMEX, and NASDAQ exchanges, where the stocks are sorted into decile portfolios based on univariate sorts. A total of 16 different datasets are used, with a varying initial starting point in time. We adjust the datasets such that the starting point for all, is July 1963 and ending December 2019. This time period is chosen in order to conform the datasets to each other, with an identical starting and ending point. This is also consistent with the starting point of previous literature investigating the topic.

Table 4.1: US datasets

#	Data and source	Abbreviation	Time period
1	Portfolios formed on Size	SIZE	7 / 1926 - 12 /2019
2	Portfolios formed on Book-to-Market	BM	7 / 1926 - 12 /2019
3	Portfolios formed on Operating Profitability	OP	7 / 1926 - 12 /2019
4	Portfolios formed on Investment	INV	7 / 1926 - 12 /2019
5	Portfolios formed on Earnings / Price	EP	7 / 1951 - 12 /2019
6	Portfolios formed on Cashflow / Price	CP	7 / 1951 - 12 /2019
7	Portfolios formed on Dividend Yield	DIV	7 / 1927 - 12 /2019
8	Portfolios formed on Momentum	MOM	7 / 1927 - 12 /2019
9	Portfolios formed on Short-Term reversal	SHORT	7 / 1926 - 12 /2019
10	Portfolios formed on Long-Term reversal	LONG	7 / 1931 - 12 /2019
11	Portfolios formed on Accruals	ACR	7 / 1963 - 12 /2019
12	Portfolios formed on Market Beta	BETA	7 / 1963 - 12 /2019
13	Portfolios formed on Net Shares issued	NSI	7 / 1963 - 12 /2019
14	Portfolios formed on Variance	VAR	7 / 1963 - 12 /2019
15	Portfolios formed on Residual Variance	RV	7 / 1963 - 12 /2019
16	Portfolios formed on 10-Industry	IND	7 / 1926 - 12 /2019

Table 4.1 contains the datasets downloaded from the online data library of Kenneth French for the US market. Here the name of the dataset, it's abbreviation and the time span is listed.

The Norwegian data formed from the Oslo stock exchange consists of 4 cross-sectional portfolios. Due to limitations of the data, we employ a different starting point than for the US portfolios. Namely, we set the common starting point in July 1981, ending December 2018.

Table 4.2: Norwegian datasets

#	DATA AND SOURCE	Abbreviation	TIME PERIOD
1	Portfolios formed on Size	SIZE	1 / 1980 - 12 /2018
2	Portfolios formed on book-to-market	BM	1/ 1981 - 12 /2018
3	Portfolios formed on Momentum	MOM	1 / 1980 - 12 /2018
4	Portfolios formed on Spread	SPREAD	1 / 1981 - 12 /2018

Table 4.2 contains the datasets downloaded from the online data library of Ødegaard. It contains the name of the dataset, it's abbreviation and the time span.

In addition to the portfolios formed on equity market returns, we employ factor data when computing the performance measures on the optimized portfolio strategies. We utilize the FFC4 factor model for both the US data and the Norwegian data. Both FFC4 models are downloaded from the respective online data library.

5. Empirical Results

This section is dedicated to presenting the empirical result. First, we present the results of the individual tests. For each dataset, 10 optimized strategies are tested against the null hypothesis, whether the performance of the active strategies on the applied portfolio compared to the naive is statistically different. We also present the overall values of the alphas and the Sharpe ratios. In order to compare performance results, the descriptive tables are divided between US and Norwegian computations and between the two performance measures respectively. Secondly, we present an overview of the correlation between the optimized strategies and the benchmark. Lastly, the main findings of the study are presented. Here we present the overall results of the SPA and WRC tests, and whether any optimized strategies beat the naive benchmark when looking at all datasets for the respective markets as a whole.

5.1 Individual test results

Tables 5.1-5.2 presents the individual performance measure values for each strategy on the US data. Beneath these values, the results of a one-sided t-test, measuring the significance of the difference between the optimized strategy and the naive are provided.

In Table 5.1, we observe a few cases where the alpha is significantly better from the naive 1/N, even expanding to a 10% significance level from the 5% used. Due to the low values of the alpha measurements, these are reported as percentages not as decimals, which is the case for the Sharpe ratio measurements. The results show that all of the active strategies do differ from the benchmark in some cases. Specifically, 22, (13.75%) of the combinations show significant evidence towards the strategies outperforming from the naive. All strategies have at least one case of significant difference, and at most 4 significant differences from a total of 16 possible. Risk parity and volatility timing are the two best performing strategies from this table, with 4 values significantly different from the naive, each. Nevertheless, this is only in 4 out of 16 datasets, corresponding to 25% of the total. Overall, we observe few significant differences, indicative of an overall conclusion of no optimized strategies outperforming the naive 1/N when employing the joint SPA and WRC tests.

Table 5.2 presents the same overview of the Sharpe ratios with US data. A major difference is that we here observe 49 (26.25%) significant values of the whole. Looking at specific strategies we similarly as in Table 5.4, observe risk parity and volatility timing as the best performing strategies. Significantly outperforming the benchmark in 12 of the 16 total datasets. The ERC strategy is close with 11 of 16 significant values higher than the naive. Together these three strategies all outperform the naive in the majority of the datasets. This indicates an overall conclusion that one or more of these three outperform the naive when performing the combined SPA/WRC tests. An important distinction of the Sharpe ratios compared to the alphas in Table 5.1, is that the Sharpe ratios are not presented in percentage. The values reported are therefore not directly comparable.

STRATEGY	SIZE	BM	OP	INV	EP	CP	DY	MOM	SHORT	LONG	ACR	BETA	NSI	VAR	RVAR	10IND
Naive	-0.121	0.070	-0.304	0.474	0.624	0.452	0.317	0.321	0.074	0.408	0.646	0.054	-0.527	-0.208	-0.428	0.848
ERC	-0.052	0.151	-0.156	0.553	0.596	0.432	0.45	0.258	0.017	0.466	0.761	0.29	-0.409	-0.208	-0.157	1.008
	(0.085)	(0.135)	**(0.001)**	(0.103)	(0.477)	(0.615)	(0.159)	(0.572)	(0.430)	(0.396)	**(0.014)**	(0.109)	(0.094)	(0.051)	(0.063)	(0.297)
MVP	0.243	-0.013	1.025	0.447	0.057	0.211	0.806	0.365	-0.214	0.726	1.27	1.222	-0.236	0.859	0.969	1.629
	(0.392)	(0.881)	**(0.003)**	(0.952)	(0.296)	(0.649)	(0.501)	(0.939)	(0.566)	(0.549)	(0.253)	(0.182)	(0.632)	(0.128)	(0.070)	(0.332)
MDEC	-0.417	-0.538	-0.829	0.135	0.741	0.243	0.136	-0.398	-1.052	-0.052	0.379	0.105	-0.08	-3.315	-3.438	0.909
	(0.541)	(0.142)	(0.118)	(0.242)	(0.777)	(0.614)	(0.712)	(0.169)	**(0.004)**	(0.319)	(0.332)	(0.913)	(0.143)	**(0.000)**	**(0.000)**	(0.924)
MV	-0.008	-0.741	-0.977	-0.546	-0.643	-0.224	-2.188	-0.479	1.445	-0.607	0.862	-0.758	2.153	-2.453	-1.959	1.641
	(0.914)	(0.425)	(0.423)	(0.255)	(0.178)	(0.507)	**(0.017)**	(0.456)	(0.145)	(0.340)	(0.795)	(0.469)	**(0.001)**	(0.064)	(0.109)	(0.622)
MDIV	-0.253	-0.280	-0.513	0.172	0.73	0.229	0.239	-0.012	-1.016	0.071	0.561	0.545	-0.012	-1.827	-1.999	1.096
	(0.768)	(0.359)	(0.490)	(0.268)	(0.784)	(0.571)	(0.883)	(0.478)	**(0.004)**	(0.393)	(0.746)	(0.362)	(0.113)	**(0.048)**	**(0.002)**	(0.690)
RE	-0.017	0.837	-0.661	0.539	0.616	-0.014	0.353	-0.492	-0.269	0.831	0.468	0.375	-0.129	-0.904	-0.959	1.415
	(0.728)	**(0.019)**	(0.197)	(0.818)	(0.978)	(0.114)	(0.072)	(0.119)	(0.305)	(0.291)	(0.500)	(0.333)	(0.226)	(0.453)	(0.164)	(0.177)
RP	-0.059	0.159	-0.145	0.557	0.603	0.445	0.434	0.335	0.06	0.468	0.762	0.276	-0.431	-0.172	-0.115	1.007
	(0.121)	(0.113)	**(0.001)**	(0.088)	(0.597)	(0.869)	(0.511)	(0.902)	(0.850)	(0.417)	**(0.015)**	(0.0819)	(0.090)	**(0.019)**	**(0.029)**	(0.175)
RRT	0.51	-0.260	-0.214	0.449	0.565	0.535	-0.194	-0.095	1.697	0.707	0.805	-0.566	1.335	-0.790	0.049	0.476
	(0.514)	(0.633)	(0.901)	(0.972)	(0.929)	(0.908)	(0.511)	(0.602)	(0.069)	(0.705)	(0.828)	(0.534)	**(0.021)**	(0.830)	(0.576)	(0.726)
TD	-0.125	0.311	-0.182	0.053	0.821	0.299	0.258	0.232	-0.594	0.350	0.663	0.134	-0.741	-0.785	-0.447	1.258
	(0.987)	(0.391)	(0.462)	**(0.007)**	(0.421)	(0.635)	(0.793)	(0.745)	**(0.004)**	(0.828)	(0.928)	(0.663)	(0.116)	(0.552)	(0.949)	(0.265)
VT	0.115	0.289	0.166	0.656	0.533	0.399	0.792	0.385	0.069	0.542	1.022	0.744	-0.221	0.617	0.559	1.375
	(0.113)	(0.214)	**(0.001)**	(0.225)	(0.516)	(0.700)	(0.068)	(0.831)	(0.980)	(0.531)	**(0.032)**	(0.113)	(0.148)	**(0.032)**	**(0.022)**	(0.200)

Table 5.1: Alpha values - US

The table reports the out-of-sample estimates for the time period of July 1963 to December 2019. Showing annualized alpha estimates in decimals for all tested strategies. Beneath the respective optimized strategies are the p-values for the outperformance of the alpha values from each strategy to that of the benchmark. All p-values are in percentage and in parentheses. p-values that are below a 5% significance level are bolded.

STRATEGY	SIZE	BM	OP	INV	EP	CP	DY	MOM	SHORT	LONG	ACR	BETA	NSI	VAR	RVAR	10IND
Naive	0.411	0.485	0.389	0.456	0.494	0.485	0.475	0.366	0.401	0.480	0.411	0.412	0.355	0.358	0.354	0.467
ERC	0.417	0.490	0.402	0.471	0.493	0.486	0.492	0.387	0.405	0.489	0.423	0.447	0.371	0.408	0.393	0.497
	(0.121)	(0.089)	**(0.000)**	**(0.000)**	(0.540)	(0.327)	**(0.0128)**	**(0.006)**	(0.228)	**(0.029)**	**(0.000)**	**(0.001)**	**(0.001)**	**(0.000)**	**(0.000)**	**(0.008)**
MVP	0.414	0.488	0.497	0.496	0.461	0.467	0.524	0.442	0.406	0.533	0.498	0.544	0.398	0.518	0.518	0.537
	(0.488)	(0.466)	**(0.001)**	(0.153)	(0.797)	(0.687)	(0.216)	**(0.035)**	(0.449)	(0.089)	**(0.022)**	(0.057)	(0.173)	**(0.035)**	**(0.014)**	(0.172)
MDEC	0.384	0.425	0.327	0.443	0.464	0.453	0.448	0.347	0.323	0.437	0.403	0.394	0.398	0.162	0.148	0.440
	(0.825)	(0.985)	(0.996)	(0.772)	(0.858)	(0.879)	(0.795)	(0.711)	(0.999)	(0.931)	(0.678)	(0.734)	**(0.017)**	(0.999)	(0.999)	(0.716)
MDIV	0.396	0.440	0.357	0.455	0.477	0.466	0.461	0.378	0.331	0.449	0.423	0.458	0.407	0.295	0.269	0.487
	(0.684)	(0.956)	(0.943)	(0.533)	(0.746)	(0.767)	(0.631)	(0.340)	(0.999)	(0.893)	(0.227)	(0.108)	**(0.010)**	(0.936)	(0.996)	(0.341)
MV	0.385	0.486	0.362	0.436	0.448	0.499	0.286	0.486	0.447	0.426	0.456	0.371	0.502	0.271	0.311	0.383
	(0.669)	(0.487)	(0.692)	(0.644)	(0.772)	(0.416)	(0.996)	(0.086)	(0.191)	(0.809)	(0.191)	(0.746)	**(0.002)**	(0.926)	(0.785)	(0.796)
RE	0.422	0.542	0.386	0.478	0.489	0.458	0.483	0.356	0.379	0.505	0.426	0.448	0.403	0.366	0.348	0.544
	(0.105)	**(0.005)**	(0.549)	(0.127)	(0.592)	(0.912)	(0.364)	(0.607)	(0.848)	(0.172)	(0.202)	(0.051)	**(0.014)**	(0.379)	(0.607)	**(0.013)**
RP	0.417	0.492	0.403	0.471	0.496	0.488	0.489	0.390	0.408	0.490	0.424	0.443	0.368	0.408	0.395	0.492
	(0.271)	**(0.038)**	**(0.000)**	**(0.000)**	(0.234)	(0.171)	**(0.003)**	**(0.002)**	(0.087)	(0.031)	**(0.000)**	**(0.001)**	**(0.001)**	**(0.000)**	**(0.000)**	**(0.004)**
RRT	0.433	0.521	0.409	0.519	0.539	0.553	0.436	0.519	0.509	0.532	0.472	0.411	0.485	0.398	0.453	0.436
	(0.353)	(0.231)	(0.337)	(0.125)	(0.176)	(0.097)	(0.750)	**(0.017)**	**(0.030)**	(0.176)	(0.110)	(0.505)	**(0.009)**	(0.260)	**(0.047)**	(0.651)
TD	0.406	0.455	0.381	0.400	0.487	0.450	0.475	0.384	0.349	0.437	0.409	0.424	0.340	0.361	0.364	0.502
	(0.580)	(0.918)	(0.752)	(0.999)	(0.656)	(0.935)	(0.489)	(0.191)	(0.999)	(0.989)	(0.570)	(0.158)	(0.953)	(0.454)	(0.286)	(0.113)
VT	0.431	0.505	0.431	0.495	0.497	0.490	0.529	0.433	0.422	0.507	0.455	0.511	0.399	0.506	0.479	0.538
	(0.148)	**(0.048)**	**(0.000)**	**(0.002)**	0.360	0.298	**(0.006)**	**(0.002)**	0.096	**(0.045)**	**(0.001)**	**(0.005)**	**(0.002)**	**(0.001)**	**(0.000)**	**(0.023)**

Table 5.2: Sharpe ratios - US

The table reports the out-of-sample estimates for the time period of July 1963 to December 2019. Showing annualized Sharpe ratios in decimals for all tested strategies. Beneath the respective optimized strategies are the p-values for the outperformance of the Sharpe ratios from each strategy to that of the benchmark. All p-values are in percentage and in parentheses. p-values that are below a 5% significance level are bolded.

Table 5.3 gives the individual Sharpe ratios measurements for the Norwegian data in decimals, and the corresponding p-values for a significant difference from the benchmark. With 4 datasets in total, the source data are less expansive than for the US. We nevertheless observe several strategies with one or more cases outperforming the naive, most notably the Maximum decorrelation and Maximum diversification with 3 (75%) cases each. In total, observe 12 (30%) where the optimized strategy outperforms the naive 1/N.

Table 5.3: Sharpe ratio - Norway

STRATEGY	MOMENTUM	BOOK	SPREAD	SIZE
Naive	0.637	0.771	0.655	1.312
ERC	0.625	0.780	0.693	1.338
	0(.912)	(0.205)	**(0.001)**	**(0.050**)
MVP	0.500	0.768	0.788	1.267
	(0.986)	(0.516)	**(0.043)**	(0.712)
MDEC	0.727	0.759	0.773	1.473
	(0.006)	(0.619)	**(0.019)**	**(0.008)**
MDIV	0.716	0.763	0.784	1.445
	(0.014)	(0.567)	**(0.015)**	**(0.022)**
MV	0.595	0.696	0.809	1.378
	(0.659)	(0.739)	(0.135)	(0.342)
RE	0.625	0.752	0.698	1.283
	(0.621)	(0.662)	(0.150)	(0.761)
RP	0.621	0.777	0.661	1.303
	(0.960)	(0.256)	(0.240)	(0.900)
RRT	0.615	0.672	0.919	1.474
	(0.646)	(0.915)	**(0.012)**	(0.029)
TD	0.632	0.785	0.885	1.523
	(0.574)	(0.330)	**(0.000)**	**(0.000)**
VT	0.581	0.789	0.0676	1.267
	(0.969)	(0.301)	(0.248)	(0.869)

The table reports annualized Sharpe ratios in percentage for all tested strategies. Beneath the respective optimized strategies are the p-values for the outperformance of the Sharpe ratios values from each strategy to that of the benchmark. All p-values are in percentage and parentheses. p-values that are below the 5 % significance level are bolded.

Table 5.4, reports the individual alpha measurements for the Norwegian data. We observe much higher alpha values compared to the US data, with values as high as 19.443% and as low as -5.779%. This highlights the difference in the overall markets where the factors in the FFC4 model alpha seemingly better capture the performance anomalies for the US data, or on the other hand, the optimized strategies might provide more abnormal market return in the Norwegian market. Similarly to the results from the Sharpe ratio, we observe the MDEC and MDIV strategies as the ones with the most significant p-values. Each, having three out of four (75%) significant. In total, observe 12 instances (30%) where the optimized strategy outperforms the naive 1/N.

Table 5.4: Alpha values - Norway

STRATEGY	MOMENTUM	BOOK	SPREAD	SIZE
Naive	-5.35	0.771	-5.258	5.416
ERC	-5.375	0.673	-4.722	5.547
	(0.863)	**(0.039)**	**(0.005)**	(0.523)
MVP	-6.05	3.137	-2.718	5.35
	(0.488)	(0.052)	**(0.035)**	(0.961)
MDEC	-3.407	1.167	-2.018	8.543
	(0.003)	(0.266)	**(0.000)**	**(0.002)**
MDIV	-3.468	1.448	-2.004	7.746
	(0.014)	(0.172)	**(0.000)**	**(0.015)**
MV	-5.303	1.583	2.013	19.443
	(0.621)	(0.652)	**(0.007)**	**(0.000)**
RE	-4.654	1.16	-4.237	5.151
	(0.621)	(0.287)	(0.152)	(0.716)
RP	-5.438	0.510	-5.302	5.055
	(0.548)	(0.128)	(0.729)	**(0.048)**
RRT	-5.779	-1.096	1.699	10.438
	(0.688)	(0.392)	**(0.000)**	**(0.000)**
TD	-5.448	0.179	-1.228	8.468
	(0.838)	(0.403)	**(0.000)**	**(0.000)**
VT	-5.6	1.421	-5.311	4.53
	(0.607)	(0.079)	(0.910)	(0.226)

The table reports annualized alphas in percentage for all tested strategies. Beneath the respective optimized strategies are the p-values for the outperformance of the alpha values from each strategy to that of the benchmark. All p-values are in percentage and in parentheses. p-values that are below the 5 % significance level are bolded.

5.2 Correlation values

In order to compare the similarity in performance for allocation alternatives tested, we calculate the correlation between them. From this, we can establish which strategies are close to identical, and which differs the most. A high correlation will affect the individual optimized strategies and the outperformance tests against the naive: formulae 3.19 and 3.22. In addition the SPA joint test is negatively affected by high correlation, where it can produce potentially misleading results in the final p-value.

The correlation is presented in Tables 5.5 and 5.6. For both the US and Norwegian markets, the results are similar with major correlation amongst all strategies. The correlation values are calculated means from all datasets in the respective markets. While not reported the corresponding p-values for the correlation test are significant at the 1% level in all cases. We observe ERC, risk parity and volatility timing, being highly correlated with each other and the naive benchmark, where they are reported close to 100% correlation at the three decimal level. Other strategies follow closely but with somewhat lower correlation.

Table 5.5: Correlation US

	NAIVE	RP	MVP	MDIV	MV	VT	ERC	RRT	MDEC	RE	TD
NAIVE	1	0.999	0.972	0.989	0.928	0.997	0.999	0.935	0.986	0.991	0.998
RP	0.999	1	0.975	0.988	0.930	0.998	0.999	0.936	0.984	0.992	0.996
MVP	0.972	0.975	1	0.964	0.918	0.982	0.975	0.929	0.954	0.977	0.967
MDIV	0.989	0.988	0.964	1	0.923	0.984	0.988	0.931	0.998	0.986	0.992
MV	0.928	0.930	0.918	0.923	1	0.931	0.930	0.936	0.915	0.923	0.935
VT	0.997	0.998	0.982	0.984	0.931	1	0.998	0.937	0.978	0.992	0.992
ERC	0.999	0.999	0.975	0.988	0.930	0.998	1	0.936	0.984	0.992	0.996
RRT	0.935	0.936	0.929	0.931	0.936	0.937	0.936	1	0.931	0.936	0.936
MDEC	0.987	0.984	0.955	0.998	0.915	0.978	0.984	0.925	1	0.982	0.989
RE	0.991	0.992	0.977	0.986	0.923	0.992	0.992	0.931	0.982	1	0.988
TD	0.997	0.995	0.966	0.991	0.935	0.992	0.995	0.936	0.989	0.987	1

Table 5.5 reports the average correlation for all the strategies in the 16 US datasets.

Table 5.6: Correlation Norway

	Naive	RP	MVP	MDIV	MV	VT	ERC	RRT	MDEC	RE	TD
NAIVE	1	0.998	0.940	0.979	0.834	0.985	0.998	0.946	0.979	0.974	0.989
Rp	0.998	1	0.953	0.978	0.824	0.992	0.999	0.943	0.976	0.981	0.916
MVP	0.940	0.952	1	0.929	0.743	0.978	0.952	0.878	0.917	0.961	0.970
MDIV	0.979	0.978	0.929	1	0.828	0.965	0.98	0.944	0.998	0.975	0.970
MV	0.834	0.8246	0.743	0.828	1		0.797	0.826	0.912	0.787	0.835
VT	0.985	0.992	0.977	0.965	0.797	1	0.992	0.927	0.957	0.983	0.967
ERC	0.998	0.999	0.952	0.981	0.826	0.991	1	0.944	0.978	0.982	0.986
RRT	0.946	0.943	0.877	0.944	0.912	0.927	0.944	1	0.945	0.924	0.935
MDEC	0.979	0.975	0.917	0.998	0.842	0.957	0.978	0.945	1	0.967	0.973
RE	0.975	0.981	0.961	0.975	0.787	0.984	0.982	0.924	0.9671	1	0.957
TD	0.989	0.986	0.917	0.971	0.835	0.967	0.936	0.935	0.973	0.957	1

Table 5.6 reports the average correlation for all the strategies in the 4 Norwegian datasets.

5.3 Results from joint test

The SPA test operates with a threshold criterion (A) for the z-statistics from each optimized alternative, we have earlier determined this to be -1.928 for the US data, and -1.884 for the Norwegian. From the Tables, A1-A4 and A10 found in the Appendix, pertaining to the SPA test, we observe several z-statistics below these threshold values. In order to correctly perform the SPA test, these will need to be removed due to the major infeasibility of the strategies. In Tables 5.7-5.8, we present the number of strategies removed for the US and Norwegian datasets. For the US data, we remove a total of 8 strategies due to across all datasets with the alpha measure, and 11 strategies across all datasets with the Sharpe measure. In the Norwegian datasets we remove 1 strategy with the Sharpe ratio and 0 with the alpha. Thus, in both markets, more strategies are removed for the Sharpe ratio performance measure.

Table 5.7: Number of strategies removes for the US datasets

Dataset	Alpha	Sharpe
Portfolios formed on Size	0	1
Portfolios formed on Book-to-Market	0	0
Portfolios formed on Operating Profitability	0	0
Portfolios formed on Investment	0	1
Portfolios formed on Earnings / Price	0	1
Portfolios formed on Cashflow / Price	1	1
Portfolios formed on Dividend Yield	1	1
Portfolios formed on Momentum	0	0
Portfolios formed on Short-Term reversal	0	0
Portfolios formed on Long-Term reversal	1	0
Portfolios formed on Accruals	0	0
Portfolios formed on Market Beta	2	2
Portfolios formed on Net Shares issued	0	0
Portfolios formed on Variance	3	3
Portfolios formed on Residual Variance	0	1
Portfolios formed on 10-Industry	0	1
Strategies removed in total	**8**	**11**

Table 5.7 contains the number of strategies removed for each of the 16 US datasets, according to the z-stat calculated from the alpha or Sharpe ratio.

Table 5.8: Number of strategies removes for the Norwegian datasets

Dataset	Alpha	Sharpe
Portfolios formed on Size	0	1
Portfolios formed on Book-to-Market	0	0
Portfolios formed on Operating Profitability	0	0
Portfolios formed on Investment	0	0
Strategies removed in total	**0**	**1**

Table 5.8 contains the number of strategies removed for each of the 4 Norwegian datasets, according to the z-stat calculated from the alpha or Sharpe ratio.

The maximum z-value from each strategy is utilized to compute the combined test p-value. A significant p-value pertains to at least one strategy outperforming the benchmark. The p-values calculated for the different performance measures are listed in Table 5.9 below. The corresponding test statistic distribution for each test are shown in Figure 5.1-5.4.

Table 5.9: Joint test computation US & Norway

US dataset		
Test	Performance measure	P-value
SPA	Alpha	0.6436
SPA	Sharpe	0.0053
WRC	Alpha	0.9155
WRC	Sharpe	0.2311
Norwegian dataset		
Test	Performance measure	P-value
SPA	Alpha	0.000
SPA	Sharpe	0.000
WRC	Alpha	0.001
WRC	Sharpe	0.638

Table 5.9 contains the different p-values calculated by the join test. The results are composed of the US and Norwegian datasets. Within each of these, the p-values for the SPA and WRC are sorted by the respective performance measures Alpha and the Sharpe ratio.

For the US data, we observe a divide between the two performance measures in the SPA test. When testing the alpha values from each strategy, the results are conclusive in that we cannot reject the null hypothesis of equality between the optimized strategies and the benchmark, with a reported p-value above a 10% significance level. When testing the Sharpe ratio values, we observe the opposite results, with a p-value beneath the 1% rejecting the H_0 and providing evidence towards at least one of the strategies beating the benchmark. Looking at the WRC test, we observe that both performance measures provide p-values above a significance level of 10%. For the alpha measure, this p-value is close to 100%. Therefore, we can not reject the null hypothesis of equality in both cases.

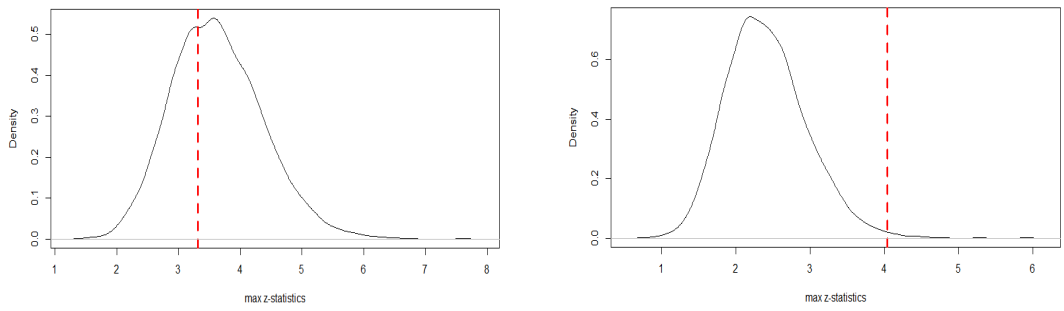
The Norwegian data is handled in the same manner as the US, and joint tests are conducted. The SPA test is conclusive for both methods at the 1% significance level. We therefore reject the null hypothesis of equally performing strategies compare to the naive. The results of the WRC test are more diverging, we observe one significant value Alpha and one non-significant Sharpe ratio.

5.4 Distribution of the maximum z-statistics

These distribution Figures 5.1-5.4 correspond to the highest z-statistics $\overline{f}$ of the respective p-values from the join test. The joint test results are found by bootstrap the data 10,000 times. The position of the red line in relation to the distribution in suggests if the test finds at least one outperforming strategy or not. The red line represents the maximum z-statistic. If the distribution is below, it suggests a p-value which is below the significant value of 0.05. The distribution of the maximum z-stat for the alpha SPA suggests that the calculated p-value is above the 0.05 significant level. The Sharpe SPA suggests that the p-value is below the 0.05 significant level. This corresponds with the results from the joint-test.

The distribution of the maximum z-statistic of the US dataset illustrates that SPA Sharpe is the only one that pertains significant joint test p-value, see Figure 5.1B. This is corresponding with the result from the joint test found in Table 5.9. The maximum z-distribution formed on alphas are listed on the left side, while the once formed on the Sharpe ratios are listed on the right side.

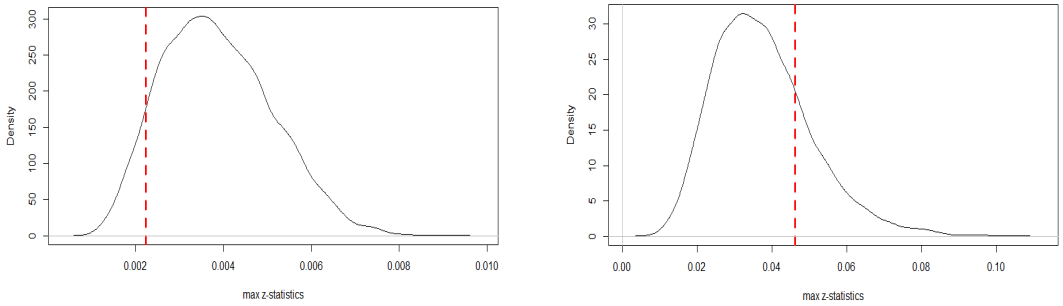
Figure 5.1: Distribution of $\overline{f}$ - SPA US



(a) Alpha SPA

(b) Sharpe ratio SPA

Figure 5.2: Distribution of $\overline{f}$ - WRC US

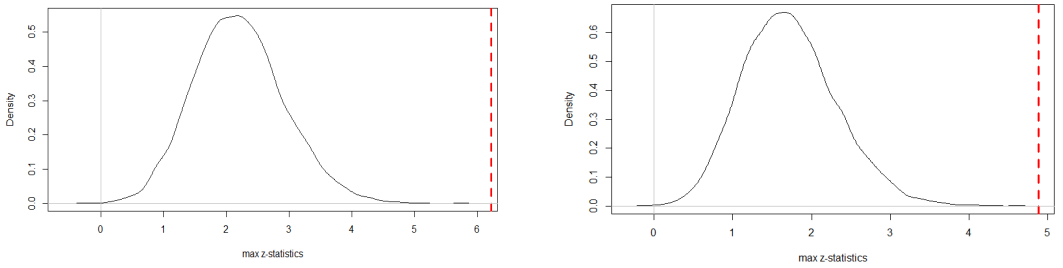


(a) Alpha WRC

(b) Sharpe ratio WRC

The distribution of the Norwegian maximum z-statistic differs from the US. We see three distribution illustrating significant joint test p-values; see Figure 5.2A, 5.2B, and 5.3a. This is corresponding with the result of the joint test in Table 5.9. The maximum z-distribution formed on alphas are listed on the left side, while the once formed on the Sharpe ratios are listed on the right side.

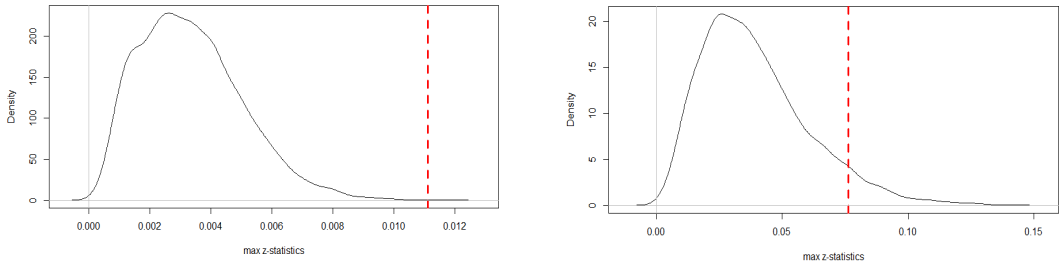
Figure 5.3: Distribution of $\overline{f}$ SPA Norway



(a) Alpha SPA

(b) Sharpe ratio SPA

Figure 5.4: Distribution of $\overline{f}$ WRC Norway



(a) Alpha WRC

(b) Sharpe ratio WRC

6. Discussion

The results of this thesis are primarily based on the computations from the two joint tests found in Table 5.9, not on the individual p-values for each strategy, as reported in Tables 5.1-5.4. Consequently, this chapter is mostly focused on discussion around these joint p-values and results.

For the US data, the alpha measure is non-significant in both tests. When testing with the Sharpe ratio measurement, the SPA test provides a significant p-value, and the WRC tests returns a non-significant value. This highlights the overall differences in our performance measure, as well as that of the underlying joint tests. The similar results for both tests with the alpha measure can partially be explained in the low alpha values observed in Table 5.1. Most of these values are found to be not only non-significant in their difference towards the naive, but also so low that they in themselves are non-significant. Therefore, any differences in the computation of the final p-values from the SPA and WRC tests will become marginalized by the overall insignificant alpha values used.

The low alphas observed in the US data is notable and may indicate that the alpha model employed in this thesis is a good fit. It expands upon the Fama-french three-factor model by including the proposed momentum factor by Carhart (Carhart, 1997). While not reported in the empirical results, the addition of the momentum factor drastically lowers the overall alphas measured.

When employing the joint tests on the Norwegian data, we observe an opposite result from the alpha estimates compared to the US data. Here we find highly significant p-values from both tests. Indicating that the best strategy, as measured by the FFC4 alpha, significantly outperforms the naive. This result can also be seen in the individual alphas measured from each optimized strategy, where they are much higher compared to the US data, and more often are found to significantly outperform the benchmark on an individual basis as seen in Table 5.4. Regarding the Sharpe ratio, both tests are

similar in their results compared to the respective tests for the US data. Where the SPA Sharpe ratio test finds significance, and the WRC Sharpe ratio does not. Consequently, this indicates that the Sharpe ratio measure produces more similar results across the two markets.

The difference in the result when estimating with the performance measures accentuates the diverging methodologies of SPA and WRC tests. The SPA test is argued as an improvement over the WRC due to the removal of irrelevant strategies and the test statistic being normalized/studentized (Hansen, 2005). Combined, these two features explain the difference in results from the joint tests we observe in Table 5.9. However, high correlation amongst the alternative strategies might skew the results of the SPA test (Hansen, 2005). As shown in Table 5.5 and Table 5.6 there is an overall high correlation between the strategies and the naive diversification, in addition the strategies are highly correlated against each other with values close to 100%.

Using the US SPA Sharpe ratio results as an example, the high correlation between the strategies becomes apparent. By studying Table 5.2, we observe that the ERC, risk parity and volatility timing strategies produce promising results, outperforming the naive 1/N in the majority of the cases. However, these strategies are all highly correlated, and as presented in Table 5.5, close to 100% We observe that the Sharpe ratios (Table 5.2) for these strategies are only marginally different. For example, when looking at the performance calculated in the dataset "OP- operating profitability", the naive 1/N produces a Sharpe ratio of 0.389 ERC of 0.402, RP of 0.403 and VT of 0.431. These differences are all however, statistically significant at the 1% level. When performing the SPA test, these high correlation values of the strategies lowers the final p-value of the test. A statistically significant outcome might therefore not be economically significant.

The high correlation is observable amongst all strategies for both performance measures and in both markets. It might partially explain the difference we observe when comparing the two joint tests for the Sharpe ratio performance measure. The WRC test does not exhibit the correlation problem and therefore produces opposite results, not rejecting the null for both markets. Interestingly we observed similar results when employing the alpha performance measure, concluding in non-significance for the US data and in significance for the Norwegian. Therefore the tests only differ in their conclusion when employing the Sharpe ratio.

The difference in the performance measures is another reason for the dissimilarity in the joint test. In the SPA test, we had to remove more strategies which were measured by Sharpe ratio than for the FFC4 alpha. For the Norwegian datasets we removed one strategy when measuring with the Sharpe ratio and none with the FFC4 alpha, and in the US removed 8 measured by alpha and 11 measured by Sharpe. While not conclusive, this indicates a discrepancy in the performance measures. Noting this, we refer to the article by Zakamulin (2017) discussed in the literature review, arguing alpha as a more expansive and encompassing measure, and that the Sharpe ratio is incomplete when explaining the source of performance gains. The Sharpe ratios might therefore not accurately measures the difference in performance for our strategies tested. A more comprehensive study of different results obtained would be interesting, highlighting the potential discrepancy in quality.

In total, there are 8 different joint tests produced in this thesis. If we consider them all together, the major divide of our results lies in the two markets analyzed, US and Norwegian data. With 3 out of 4 tests concluding in non-significance for the US datasets, and 3 out of 4 concluding in significance for the Norwegian datasets. Because of the high correlation amongst the alternate strategies, and the negative effect it has on the SPA test, the results from the WRC appear more valid. This is despite SPA being proposed as an improvement to the WRC test. If we then compare our WRC results to the findings of DeMiguel et al. (2009), that overall an optimized strategy does not outperform the naive. We reach the same conclusion for our US datasets when considering the results from both performance measures. Interestingly our results differ somewhat when considering the Norwegian datasets, indicating that the Norwegian market behaves/operate differently than the US, producing opposite results with the alpha measure. These differences may be present due to the market efficiency varying across the markets and time period. The US market might be argued as more efficient and "competitive" than the Norwegian.

A similar picture can be seen by comparing the articles by Hsu et al. (2018) and Yang et al. (2018) to our findings. Both articles preformed joint tests with the Sharpe ratio and alpha measure. These articles found little evidence towards any active strategies significantly outperforming the naive. This is similar to our findings for the US market, but opposite to our results for the Norwegian.

In this thesis we only investigate whether there exists one strategy which outperforms

the naive benchmark. This is due to the tests employing only the maximum z-stat from all available in their computation. While the joint tests on the US data are primarily insignificant in their results, the Norwegian results suggests overall significance. We observe that two strategies stand out for the Norwegian data; maximum-decorrelation and most diversified. This suggests that more than one strategy might outperform. Observing the correlation Table 5.6, the two strategies have a somewhat high correlation towards the naive compared to all 11 alternatives, and the overall correlation value is closing on 100%. Thus, even though the individual p-values from Table 5.1-5.2 and the final joint test p-values from Table 5.9 are statistically significant, they might not be economically significant, any difference being marginal in a market setting.

When estimating the excess returns of all our allocation strategies, we do not take into account transaction costs that would be present in reality. For the optimized strategies that require constant rebalancing of weights in order to follow their weight rules, such transaction costs would be substantially higher than for the naive benchmark. Therefore, any conclusion of such strategies outperforming the naive, as seen in the Norwegian data, would be suspect. Depending on the degree of transaction costs, this conclusion could be reversed overall. It would be relevant to further include such costs and study the effect on the allocation alternatives in further studies.

7. Conclusion

The paper of DeMiguel et al. (2009) started a heated debate on the viability of optimized trading strategies. They show that none of the tested strategies outperforms the naive diversification. Several authors later claim to defend the viability of optimized strategies (Kriztman et al. (2010), Tu & Zhou (2011), Kirby & Ostdiek (2012) and Banerjee & Hung (2013)). However, White (2000) showed that the superior performance found might not be due to actual performance, but rather pure chance, due to data-mining bias. Furthermore, Zakamulin (2017) criticizes the use of the Sharpe ratio as the single performance measure. Where any measured superior performance of the optimized strategies instead might come from one or more profit anomalies, and not from real genuine outperformance.

More recently, two papers by Hsu et al. (2018) and Yang et al. (2018) attempts to account for data mining bias by including the WRC (White, 2000) and SPA (Hansen, 2005) joint tests. They find that there is little evidence of any optimized strategies outperforming the naive diversification. In this thesis, we attempt to replicate their results, and expand the study with additional datasets, and to the Norwegian market.

We compare the performance of 10 active trading strategies against the naive diversification, accounting for potential data-snooping bias (White, 2000). We seek to address this by using two advanced econometric methods called SPA and WRC test. However, the SPA test is weakened by a high correlation amongst our optimized strategies. Therefore, the results from the WRC tests are overall more valid. Additionally, alpha has been included as a performance measure in response to Zakamulin's (2017) criticism of employing Sharpe ratio as the single performance measure. Due to argued weaknesses of the Sharpe-ratio performance measure, we base the overall conclusion of the thesis on the alpha performance measure.

Our results from the US market shows little empirical evidence towards any active trading strategies significantly outperforming the naive diversification. This is in agreement with the overall findings of DeMiguel et al. (2009), and provide counter-evidence towards the criticism of Kriztman et al. (2010), Tu & Zhou (2011), Kirby & Ostdiek (2012) and Banerjee & Hung (2013).

In contrast, the Norwegian results show that most tests provide empirical evidence of the best active trading strategy significantly outperforming the naive diversification. The Norwegian results are, however, difficult to compare to previous findings due to limited academic studies in the market. Furthermore, the datasets and data period analyzed is shorter than for the US results.

Overall, we find that there is little empirical evidence of the active trading strategies significantly outperforming the naive diversification in the US data. In the Norwegian data, we provide empirical evidence of at least one of the 10 active trading strategies significantly outperforming the naive diversification, when measuring with the alpha measure.

A. References and appendix

A.1 References

Adame - Garcia, Victor, Fernandez Rodriguez, Fernando, & Sosvilla Rivero, Simon. (2017). Resolution of optimization problems and construction of efficient portfolios: An application to the Euro Stoxx 50 index. *IREA*, 201702.

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2012). Leverage aversion and risk parity. *Financial Analysts Journal*, 68(1), 47-59.

Ardia, D., Bolliger, G., Boudt, K.,Gagnon-Fleury, J.-P. (2017). The impact of covariance misspecification in risk-based portfolios. *Annuals of Operation Research*, 254(1), 1-16.

Blitz, D. C. & Van Vliet, P. (2007). The volatility effect: Lower risk without lower return. *Journal of Portfolio Management*, 102-113.

Blitz, D. (2016). The value of low volatility. *Journal of Portfolio Management*, 42(3), 94-100.

Carhart, M. (1997). On persistence in mutual fund performance.*Journal of Finance*, 52(1), 57-82.

Chaves, D., Hsu, J., Li, F., & Shakernia, O. (2011). Risk parity portfolio vs. other asset allocation heuristic portfolios. *Journal of Investing*, 20(1), 108-118.

Choueifaty, Y. (2006) Methods and systems for providing an anti-benchmark portfolio. *USPTO* , 60, 816,276.

Choueifaty, Y, & Coignard, Y. (2008) Toward maximum diversification. *Journal of Portfolio Management*, 34(4), 40-51.

Choueifaty, Y., Froidure, T., Reynier, J. (2011). Properties of the most diversified portfolio. *Journal of Investment Strategies*, 2(2), .49-70.

Christoffersen, P., Errunza, V., Jacobs, K., & Langlois, H. (2012). Is the potential for international diversification disappearing? A dynamic copula approach. *Review of Financial Studies*, 25(12), 3711-3751.

Clarke, R., DeSilva, H., & Thorley, S. (2006). Minimum-variance portfolios in the US equity market. *Journal of Portfolio Management*, 33(1), 10-25.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/ N portfolio strategy? *Review of Financial Studies*, 22(5), 1915-1953.

Fama, E., & Blume, M. (1966). Filter rules and stock-market trading. *Journal of Business*, 39(1), 226-241.

Fama, E. F. & French, K. R. (1993). Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* , 33(1), 3-56.

Fleming, J., Kirby, C., & Ostdiek, B. (2002). The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, 67(3), 473-509.

Fleming, J., Kirby, C., & Ostdiek, B. (2001). The economic value of volatility timing. *Journal of Finance*, 56(1), 329-352.

Goltz, F., Sivasubramanian, S. (2018). Scientific beta maximum decorrelation indices. *ERI scientific beta publication.*

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

Haugen, R. & Baker, N. (1991). The efficient market inefficiency of capitalization-weighted stock portfolios. *Journal of Portfolio Management*, 17(3), 35-40.

Hsu, P.-H., Han, Q., Wu, W., & Cao, Z. (2018). Asset allocation strategies, data snooping, and the 1 / N rule. *Journal of Banking and Finance*, 97(1), 257-269.

Kirby, C., & Ostdiek, B. (2012). It's all in the timing: Simple active portfolio strategies that outperform naive diversification. *Journal of Financial and Quantitative Analysis*, 47(2), 437-467.

Kan, R., & Zhou, G. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3), 621-656.

Kritzman, M., Page, S., & Turkington, D. (2010). In defense of optimization: The fallacy of 1/ N. *Financial Analysts Journal*, 66(2), 31-39.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock-portfolios and capital budgets.*Review of Economics and Statistics*, 47(1), 13-37.

Maillard, S. and Roncalli, T. and Teiletche, J.(2008) The properties of equally weighted risk contribution portfolios. *Journal of Portfolio Management*, 36(4), 60-70.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77-91.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* 34(4), 768-783.

Neurich, T. (2008). Alternative indexing with the MSCI World Index, *SSRN*

Michaud, R. (1989). The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1), 31-42.

Nelsen, R., Quesada-Molina, J., Rodriguez-Lallena, J., & Ubeda-Flores, M. (2001). Distribution functions of copulas: A class of bivariate probability integral transforms. *Statistics and Probability Letters*, 54(3), 277-282.

Qian, E. (2011). Risk parity portfolios: Effcient portfolios through true diversifcation, *Panagora Asset Management*, 20(1), 119-127.

Scherer, B. (2011). A note on the returns from minimum variance investing. *Journal of Empirical Finance*, 18(4), 652-660.

Schmidt, R., & Stadtmuller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian Journal of Statistics*, 33(2), 307-335.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3), 425-442.

Sharpe, W. (1966). Mutual fund performance. *Journal of Business*, 39(1), 119-138.

Tu, J., & Zhou, G. (2011). Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, 99(1), 204-215.

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

Yang, j., Cao, Z., Han, Q., Wang, Q., (2018). Tactical asset allocation on technical trading rules and data snooping. *Pacific-basin Finance Journal*, 57(C)

Zakamulin, V. (2017). Superiority of optimized portfolios to naive diversification: Fact or fiction? *Finance Research Letters*, 22, 122-128.

A.2 SPA alpha - US

The following Tables A.1 - A.8 contains maximum z-statistics calculated from the US datasets, using the SPA method. A.1 - A.4 contains the statistics calculated from alphas, and A.5 - A.8 contains the statistics calculated from the Sharpe ratios. The strategies are listed on the left-hand side, while the maximum z-statistic and its p-value for the dataset are listed on the top of the table. The datasts follows the notation given in table 4.1

Table A.1: SPA alpha US data 1:4

Strategy	z-OP	p-OP	z-SIZE	p-SIZE	z-BETA	p-BETA	z-BM	p-BM
RP	4.306	**0.000**	2.082	**0.019**	3.263	**0.001**	2.571	**0.005**
MVP	3.560	**0.000**	1.460	0.072	2.188	**0.015**	0.697	0.243
MDIV	-0.597	0.725	-0.192	0.576	1.219	0.112	-0.468	0.680
MV	0.356	0.361	0.359	0.360	0.544	0.293	0.272	0.393
VT	4.257	**0.000**	2.244	**0.013**	2.946	**0.002**	2.262	**0.012**
ERC	4.274	**0.000**	2.127	**0.017**	3.129	**0.001**	2.475	**0.007**
RRT	1.148	0.126	1.244	0.107	0.689	0.246	0.922	0.178
MDEC	-1.781	0.962	-0.584	0.720	-0.557	0.711	-1.212	0.887
RE	-0.294	0.616	0.545	0.293	2.010	**0.022**	3.097	**0.001**
TD	0.953	0.171	0.059	0.477	1.155	0.124	0.366	0.357

Table A.2: SPA alpha US data 2:4

Strategy	z-DIV	p-DIV	z-VAR	p-VAR	z-INV	p-INV	z-CP	p-CP
RP	1.442	**0.075**	3.863	**0.000**	2.292	**0.011**	-0.176	0.570
MVP	0.194	0.423	2.379	**0.009**	0.049	0.481	-0.706	0.760
MDIV	-1.128	0.870	-1.422	0.922	-0.510	0.695	-0.362	0.641
MV	-1.563	0.941	-0.113	0.545	0.451	0.326	0.508	0.306
VT	1.393	0.082	3.452	**0.000**	1.585	0.057	-0.521	0.699
ERC	0.848	0.198	3.470	**0.000**	2.266	**0.012**	-0.536	0.704
RRT	0.096	0.462	1.545	0.061	1.593	0.056	1.241	0.108
MDEC	-1.293	0.902	-4.825	1.000	-0.587	0.721	-0.425	0.665
RE	0.318	0.375	0.392	0.348	0.623	0.267	-1.133	0.871
TD	-0.864	0.806	-0.030	0.512	-3.186	0.999	-0.338	0.632

Table A.3: SPA alpha US data 3:4

Strategy	z-EP	p-EP	z-NSI	p-NSI	z-ACR	p-ACR	z-RV	p-RV
RP	-0.191	0.576	2.756	**0.003**	2.995	**0.001**	3.968	**0.000**
MVP	-0.926	0.823	0.830	0.203	1.887	**0.030**	3.012	**0.001**
MDIV	0.137	0.446	2.176	**0.015**	0.739	0.230	-2.214	0.986
MV	-0.481	0.685	4.194	**0.000**	1.859	**0.032**	0.038	0.485
VT	-0.416	0.661	2.417	**0.008**	2.744	**0.003**	3.890	**0.000**
ERC	-0.649	0.742	2.752	**0.003**	3.136	**0.001**	3.699	**0.000**
RRT	0.742	0.229	3.049	**0.001**	1.726	**0.042**	2.430	**0.008**
MDEC	0.059	0.477	1.969	**0.025**	-0.154	0.561	-5.733	1.000
RE	-0.131	0.552	1.973	**0.024**	0.266	0.395	-0.505	0.693
TD	0.554	0.290	-2.115	0.983	0.403	0.344	0.784	0.217

Table A.4: SPA alpha US data 4:4

n	z-MOM	p-MOM	z-SHORT	p-SHORT	z-LONG	p-LONG	z-IND	p-IND
RP	2.876	**0.002**	1.422	**0.078**	1.702	**0.045**	2.459	**0.007**
MVP	2.028	**0.021**	0.541	0.294	1.436	0.076	1.477	0.070
MDIV	0.888	0.188	-2.897	0.998	-0.804	0.789	0.974	0.165
MV	3.123	**0.001**	1.476	0.070	-0.266	0.605	2.009	**0.022**
VT	2.897	**0.002**	1.470	0.071	1.522	0.064	1.988	**0.024**
ERC	2.624	**0.004**	0.818	0.207	1.722	**0.043**	2.221	**0.013**
RRT	3.237	**0.001**	2.640	**0.004**	1.374	0.085	1.577	0.058
MDEC	-0.145	0.557	-3.242	0.999	-1.202	0.885	0.366	0.357
RE	0.309	0.379	-0.584	0.720	1.215	0.112	2.573	**0.005**
TD	2.007	**0.023**	-3.127	0.999	-0.832	0.797	1.904	**0.029**

A.3 SPA Sharpe US

Table A.5: SPA Sharpe US data 1:4

Strategy	z-OP	p-OP	z-SIZE	p-SIZE	z-BETA	p-BETA	z-BM	p-BM
RP	4.041	**0.000**	1.237	0.108	3.052	**0.001**	1.778	**0.038**
MVP	3.244	**0.001**	0.039	0.484	1.576	0.057	0.083	0.467
MDIV	-1.578	0.943	-0.429	0.666	1.237	0.108	-1.703	0.956
MV	-0.500	0.692	-0.395	0.654	-0.662	0.746	0.021	0.492
VT	3.855	**0.000**	1.030	0.151	2.584	**0.021**	1.666	**0.021**
ERC	3.947	**0.000**	1.159	0.123	3.029	**0.001**	1.342	0.090
RRT	0.422	0.337	0.376	0.354	-0.013	0.505	0.732	0.232
MDEC	-2.693	0.996	-0.864	0.806	-0.625	0.734	-2.179	0.985
RE	-0.123	0.549	0.580	0.281	1.631	0.051	2.582	**0.021**
TD	-0.681	0.752	-0.175	0.569	1.004	0.158	-1.391	0.918

Table A.6: SPA Sharpe US data 2:4

Strategy	z-DIV	p-DIV	z-VAR	p-VAR	z-INV	p-INV	z-CP	p-CP
RP	2.699	**0.003**	3.717	**0.000**	3.464	**0.000**	0.949	0.171
MVP	0.785	0.216	1.811	**0.035**	1.022	0.153	-0.489	0.687
MDIV	-0.335	0.631	-1.520	0.936	-0.082	0.533	-0.729	0.767
MV	-2.643	0.996	-1.448	0.926	-0.369	0.644	0.211	0.416
VT	2.478	**0.007**	3.086	**0.001**	2.815	**0.002**	0.530	0.298
ERC	2.233	**0.013**	3.459	**0.000**	3.479	**0.000**	0.448	0.327
RRT	-0.676	0.750	0.643	0.260	1.152	0.125	1.299	0.097
MDEC	-0.825	0.795	-4.808	1.000	-0.745	0.772	-1.170	0.879
RE	0.347	0.364	0.307	0.379	1.141	0.127	-0.308	0.621
TD	0.027	0.489	0.116	0.454	-4.282	1.000	-1.515	0.935

Table A.7: SPA Sharpe US data 3:4

Strategy	z-EP	p-EP	z-NSI	p-NSI	z-ACR	p-ACR	z-RV	p-RV
RP	0.724	0.234	3.252	**0.001**	3.429	**0.000**	3.715	**0.000**
MVP	-0.832	0.797	0.941	0.173	2.012	**0.022**	2.190	**0.014**
MDIV	-0.662	0.746	2.326	**0.010**	0.747	0.227	-2.637	0.996
MV	-0.747	0.772	2.874	**0.002**	0.875	0.191	-0.790	0.785
VT	0.358	0.360	2.850	**0.002**	3.161	**0.001**	3.352	**0.000**
ERC	-0.102	0.540	3.254	**0.001**	3.525	**0.000**	3.520	**0.000**
RRT	0.930	0.176	2.349	**0.009**	1.226	0.110	1.675	**0.047**
MDEC	-1.071	0.858	2.110	**0.017**	-0.461	0.678	-5.814	1.000
RE	-0.232	0.592	2.186	**0.014**	0.834	0.202	-0.271	0.607
TD	-0.402	0.656	-1.671	0.953	-0.177	0.570	0.566	0.286

Table A.8: SPA Sharpe US data 4:4

Strategy	z-MOM	p-MOM	z-SHORT	p-SHORT	z-LONG	p-LONG	z-IND	p-IND
RP	2.851	**0.002**	1.360	0.087	1.872	**0.031**	2.653	**0.004**
MVP	1.813	**0.035**	0.129	0.449	1.347	0.089	0.946	0.172
MDIV	0.412	0.340	-3.034	0.999	-1.242	0.893	0.410	0.341
MV	1.367	0.086	0.875	0.191	-0.877	0.810	-0.829	0.796
VT	2.790	**0.003**	1.302	0.096	1.691	0.045	1.988	**0.023**
ERC	2.538	**0.006**	0.744	0.228	1.893	**0.029**	2.390	**0.008**
RRT	2.131	**0.017**	1.875	**0.030**	0.929	0.176	-0.387	0.651
MDEC	-0.556	0.711	-3.244	0.999	-1.481	0.931	-0.571	0.716
RE	-0.272	0.607	-1.029	0.848	0.947	0.172	2.234	**0.013**
TD	0.875	0.191	-3.441	1.000	-2.307	0.989	1.212	0.113

A.4 SPA Sharpe and alhpa - Norway

The following Tables A.9 and a.10 contains maximum z-statistics calculated from the US datasets, using the SPA method. A.10 contains the statistics calculated from alpha, and A.9 contains the statistics calculated from the Sharpe ratios. The strategies are listed on the left-hand side, while the maximum z-statistic and it's p-value for the dataset are listed on the top of the table. The datasts follows the notation given in table 4.2

Table A.9: SPA Sharpe Norwegian data 1:1

Strategy	z-MOM	p-MOM	z-SIZE	p-SIZE	z-SPREAD	p-SPREAD	z-BM	p-BM
RP	-1.756	0.960	-1.285	0.901	0.705	0.240	0.655	0.256
MVP	-2.218	0.987	-0.558	0.712	1.712	0.043	-0.042	0.517
MDIV	2.172	0.015	2.009	0.022	2.158	0.015	-0.171	0.568
MV	-0.411	0.659	0.405	0.343	1.103	0.135	-0.643	0.740
VT	-1.870	0.969	-1.124	0.870	0.680	0.248	0.521	0.301
ERC	-1.357	0.913	1.643	0.050	3.015	0.001	0.824	0.205
RRT	-0.375	0.646	1.884	0.030	2.246	0.012	-1.375	0.915
MDEC	2.472	0.007	2.427	0.008	2.067	0.019	-0.304	0.619
RE	-0.310	0.622	-0.709	0.761	1.034	0.151	-0.419	0.662
TD	-0.187	0.574	4.635	0.000	4.879	0.000	0.439	0.330

Table A.10: SPA Alpha Norwegian data 1:1

Strategy	z-MOM	p-MOM	z-SIZE	p-SIZE	z-SPREAD	p-SPREAD	z-BM	p-BM
RP	0.024	0.491	-1.870	0.969	0.266	0.395	1.872	0.031
MVP	-0.277	0.609	0.042	0.483	2.381	0.009	2.287	0.011
MDIV	2.537	0.006	2.567	0.005	3.489	0.000	1.680	0.047
MV	0.134	0.447	3.830	0.000	2.839	0.002	0.506	0.307
VT	0.099	0.461	-1.091	0.862	0.357	0.361	2.052	0.020
ERC	0.240	0.405	0.700	0.242	3.182	0.001	2.451	0.007
RRT	0.031	0.488	3.489	0.000	4.021	0.000	-0.499	0.691
MDEC	2.399	0.008	3.258	0.001	3.380	0.000	1.322	0.093
RE	1.065	0.144	-0.347	0.636	1.536	0.063	1.381	0.084
TD	-0.537	0.704	5.395	0.000	6.215	0.000	1.584	0.057

A.5 Correlation of the active strategies towards the banchmark

Table A.11: Correlation of optimized strategies against the naive, US.

Dataset	Mean	Median	Maximum	Minimum	Most correlated	Least correlated
Size	0.965	0.983	0.999	0.905	ERC/RP	GVM
BM	0.974	0.985	0.999	0.900	ERC/RP	MV
OP	0.980	0.991	1	0.928	ERC/RP	MV
INV	0.978	0.992	1	0.922	ERC/RP	MV
EP	0.975	0.986	1	0.904	ERC/RP	MV
CP	0.972	0.986	1	0.892	ERC/RP	MV
DY	0.958	0.978	0.999	0.871	RP	MV
MOM	0.952	0.975	0.998	0.804	ERC/RP	MV
SHORT	0.976	0.988	0.999	0.916	ERC/RP	RRT
LONG	0.972	0.984	0.999	0.904	ERC/RP	MV
ACR	0.979	0.993	0.999	0.932	ERC/RP	MV
BETA	0.951	0.972	0.997	0.822	ERC/RP	MVP
NSI	0.976	0.989	1	0.923	ERC/RP	RRT
VAR	0.943	0.957	0.956	0.803	ERC/RP	MVP
RVAR	0.958	0.971	0.999	0.866	ERC/RP	MVP
10IND	0.922	0.957	0.998	0.734	ERC/RP	MV

Table A.11 reports the correlation of the optimized strategies against the naive for the 16 datasets employed. Column 2-4 reports the mean, median, maximum and minimum correlation values respectively. Lastly column 5-6 report the highest and lowest correlating strategies. All values are in percentages.

Table A.12: Correlation of optimized strategies against the naive, Norwegian

Dataset	Mean	Median	Maximum	Minimum	Most correlated	Least correlated
Spread	0.920	0.958	0.999	0.691	ERC/RP	MV
Size	0.926	0.958	0.998	0.672	ERC/RP	MV
MOM	0.963	0.980	0.999	0.835	ERC/RP	MV
BTM	0.949	0.972	0.999	0.790	ERC/RP	MV

Table A.12 reports the correlation of the optimized strategies against the naive for the 4 employed. Column 2-4 reports the mean, median, maximum and minimum correlation values respectively. Lastly column 5-6 report the highest and lowest correlating strategies. All values are in percentages.

A.6 Reflection notes

Reflection note, Andreas Wehus.

In a few weeks I will deliver my masters thesis and complete my master's degree in economics and administration, with a specialisation within finance. Alongside delivering my masters, the university of Agder have asked me to write this reflection note regarding my thesis and the knowledge I've obtained during my studies. This reflection note will begin with a recap of my study, followed by some key concept I believe played a prominent role, and lastly, I will talk about my master's.

During my study at the university of Agder I have completed a bachelor's in business administration (accounting) and a master's as civil economist, with specialisation in finance The subjects I took during my bachelor focused on subjects regarding law, accounting, organization theory and finance. Combined, these subjects have widened my perspective on the economy works both in a microeconomics and macroeconomics sense. During this period, I had certain topics which I found intriguing and lead to my choice of studying finance as a masters. These subjects surrounded quantitative analysis surrounding businesses, projects stocks etc. It started with theoretical mathematics, given a solid foundation. This was later put into use with quantitative finance, computational finance, econometrics and advanced econometrics continued this foundation. Quantitative finance heavily focused on the calculation of key values of stocks such as expected return, standard deviation and applied these to form portfolios and measure its performance. While this was done by pen and paper, computational finance used the statistical program R to quantify these calculations and making it possible to learn how to research on a much wider scale, using real numbers and historical data as a simulation. Econometrics and advanced econometric also applied statistical computer programming to apply methods on regression using STATA. This helped us learn how a dependent variable change in accordance with the size of the coefficient when an independent variable change. These are the subjects I believe made it possible for me to write such a quantitative master thesis, specially since its rellies heavily on my competence within the statistical programs.

During my five years of studying I have acquired a lot of knowledge regarding economics. I've learned that economics is not strictly about numbers, but also organizations, international relationships, ethics, innovation, marketing, management

and the list goes on. The combination of these subjects changes the perception one has one the economy. One acronym which I learned on my first year, which I still remember very well is the acronym CSR. CSR stands for corporate social responsibility and is a theory that the corporations have a social responsibility towards the social. Social being the community, world, people etc. The more I thought about it, the more I liked the idea of corporation having a responsibility to improve the world around it. Of course, this is not a demand, but a proposal. A business who is performing well financially could use some of it profit and put it to good use without financial gains. Examples of this could be to donate money, sponsor local sports team or invest more environmental solution for your business.

Alongside corporate social responsibility, the understanding of business across boarders was also an interesting subject. How businesses expand to other countries, outsources, imports or exports goods. Some businesses choose to move some of their production to other countries to save money on labour and taxes. This a regular occurrence and a lot of businesses are international. Preparing us for a world where businesses are international or expands to other countries, organizational theory helps us understand that there exist different entry barriers making expanding across boarders challenging. When studying we learned how to deal with these international businesses when it comes to accounting. Using IFRS, International Financial Reporting Standard, which is a common set of rules that makes the financial statements consistent, transparent and usable around the world. This makes it easier for companies to compare financial statements from companies operating in different countries.

The world is constantly changing, and how one adapts to these changes can play a major difference in the survival of businesses. Inventors, given that they are successful, can be the first adaptors and reap the benefits and outcompeting their competitors, while the last to adapt may face bankruptcy. I have learned that paying attention to how the market operate is key to surviving. Starting projects, gathering founding and creating a financial plan to discover the net present value of investing is something I've learned throughout my studies. After learning different methods of planning, funding and creating a financial plan, our knowledge was tested. We were tasked to invent a product, show how we will fund it, creating a financial plan and estimating the worth of the project. This was both challenging and exciting. Testing your knowledge in a simulated scenario is a lot different. Here you must combine all you have learned to

create something which was more challenging than separating the process and only doing the individual parts alone. Reminiscing back to my studies, I feel like I have learned a great lot about internationalisation, innovation and responsibility, which I believe will affect me in my future career.

My master's thesis is about portfolio optimization. Portfolio optimization is the process of finding the optimal portfolio combination out of a given set. Financial researchers have discovered different strategies which tells the investor how to best allocate your wealth into different portfolios to best maximize returns. These strategies are referred to as "active trading strategies", which require one to look at historical data and calculate the covariance-variance and expected returns. This method of portfolio optimization is called the modern portfolio theory (MTP) and was created by Markowitz (1952). DeMiguel et.al (DeMiguel 2009) wanted to test the performance of an active trading strategy against a naive-diversification. A naive diversification divides your wealth equally amongst all investment object, giving the expression 1/N. Doing so, he discovered that none of the active trading strategies could significantly outperform the naIve diversification.

This started a heated debate and financial researchers tried to provide counterevidence to their findings. See; Kirby Ostediek (2011) and Krietzman, Page & Turkington (2010) for examples. Later studies have shown that the results of these studies are not significant, since the methods they applied makes the result invalid. Bad usage of performance measure, few datasets and not accounting for data-snooping bias was mentioned in the critique to Krietzman et al. (2010), see: Zakamulin (2017)

Our research question is "Does any active trading strategy outperform the naIve diversification", where we improve the method of previous literature to give a more accurate result than previous researchers have. To figure this out, we apply 10 active trading strategies as well as the naIve, to 16 different datasets. The data generated from this is then loaded into a joint-test-program. Here, we apply two different test and bootstrap the data, resampling it 10.000 times before we jointly calculate if there are significant differences between the naive diversification and the active strategies. These joint in accordance to the method of Hansen (2005) and White's (2000) SPA test and WRC test, respectively. This is also done with Norwegian datasets, but it's however limited due to only four datasets being available on Norwegian dataset.

Summarized, our results mostly support that no optimized strategy outperform the

naive in the US market. For the Norwegian market we find the opposite results, with most tests pointing towards one or more strategy outperforming the naive.

Reflection note, Sindre Tjetland.

As a part of our final master thesis, The School of Business and Law at University of Agder requires that we write an individual reflection note with a brief summary of our thesis, alongside how our work relates to three key concepts of; International, innovative and responsible

In this thesis I have alongside my fellow student Andreas Wehus explored the topic of portfolio optimization. We have attempted to answer the question whether an optimized statistical strategy of asset allocation, can beat a naive strategy of equal distribution of funds in the assets.

In order to expand upon previous literature, we have attempted to incorporate additional statistical methods in our econometric estimation. Traditionally, the success of portfolio optimization has been measured in Sharpe-ratio (Sharpe, W. 1966) While this performance measure is employed in this thesis for ease of comparison, we additionally include a second performance measure; Alpha. Specifically we refer to our supervisor's paper on the topic, Zakamulin(2017) Where he discusses the weaknesses of the Sharpe-measure, and compares it to a form of alpha measure developed by Fama & French (1993) We employ this FF3 alpha in order to better capture abnormal performance in our optimized strategies. Overall, it is theorized that this makes for a superior performance measure.

Employing the FFC4 alpha measure, we further expand upon previous literature by incorporating a joint test to test all our individual alpha measures from each optimized strategy and their respective datasets. We employ two such join tests in this thesis, the original novel WRC test developed by White (2000) and its expansion the SPA test developed by Hansen (2005)

These tests are then performed in conjunction with a form of monte Carlo analysis, where we bootstrap the excess return data from the optimized strategies.

We perform these calculations for both US and Norwegian data. Our findings potentially highlight an overall difference of the underlying trading rules in the two markets. For the US datasets there is a majority evidence of no optimized strategy

beating the naive allocation method, especially when considering the argued alpha performance measure. For the Norwegian data we conclude in the opposite. Here most of the joint test results and especially the alpha provide evidence towards one or more of the optimized strategies beating the naive.

A potential weakness of our study is the difficulty of comparing the two distinct markets. We employ datasets constructed on US and Norwegian returns, formed on different criterions. There is however a difference in datasets available for each market, where we employ 16 datasets in the US, and only 4 in the Norwegian market. This weakens the overall significance of the Norwegian results and taints an attempted comparison.

I will now attempt to cover how our thesis relates to the three key concepts. As a part of our 2-year masters, all but one course has been taught in English, and most of our professors /associate professors teaching has been of foreign stature. Especially two courses have contributed to our master's thesis: Advanced econometrics (1 and 2) and Computational Finance and Portfolio Management.

Both courses were focused on internationalized trends when teaching and made for a natural starting point for our masterâ^{TM}s thesis. In our thesis we employ American data alongside Norwegian for our estimations and findings. This is done in order for our findings to be relevant not only in Norway, but also in the US where the results can more easily be relevant for the rest of the world. Overall, our thesis is majorly influence by previous American research. Many of the optimization strategies and further methodology has been developed by American professors / academics. These tests and methodologies have become increasingly relevant for the international stage beyond the direct comparison to US data. The global trade has become increasingly interlocked in the last century, and the trend is only increasing. Capital markets and functions are interlocking and becoming co-dependent on each other. This makes an analysis employed on US or Norwegian data relevant for additional markets around the world, which behave increasingly similar. Our comparison also highlights how far this similarity has come; were we observe dissimilar results from our two markets.

Innovation is a concept which all academic papers attempt to incorporate. We attempt to expand upon previous literature by incorporating additional tests and measure. Additionally, we have attempted to perform a comparison between the Norwegian and American markets. Most significantly we include the WRC and SPA joint tests which

have not been performed on such a large section of datasets for the US market, and to our knowledge never in the Norwegian market. In addition, we have developed an understanding of modern statistical programs such as STATA and especially R. These programs lay at the forefront of econometric analysis, simplifying estimations which previously would have taken significantly more work.

Lastly, I will explore upon how the concept of responsibility has been present in our work on the master's thesis. Portfolio optimization is the process of optimizing a portfolio depending on a set of assets. In its traditional form it does not consider the specific characteristics of the assets in question. Only the statistical properties. This raises the dilemma of the potential to invest in companies which promotes irresponsible actions, whether these should be regarding climate emissions, child labor, or promoting warfare. It is possible to construct portfolios which takes ethical dilemmas and responsibility into account, and such a portfolio would potentially benefit from the increased sustainability as a result.

Furthermore, the naÃ¯ve diversification strategy is considered cheaper to implement compared to following an optimized strategy rule. This rule suffers increased transaction costs as it is required to constantly estimate and rebalancing portfolio weights. It can therefore be argued that using portfolio optimization does not add any value, producing unnecessary and irresponsible additional costs.

Overall, I feel this thesis majorly incorporates the topics of international and innovative. It is conducted with both US and Norwegian data, as well as incorporating the newest scientific literature on the topic. The thesis could expand more upon the concept of responsibility with for example including datasets which specifically exclude companies regarded as ethically questionable. While this would have been ideal, such datasets does not currently exist and would have to be constructed from scratch, which lies outside the scope of this study. Both me and my writing partner is gracious towards both our supervisor and The School of Business and Law at UiA for helping and motivating us to complete this thesis.