Optimized Portfolios vs Naive Diversification: A re-examination using tests free of data-snooping

EIVIND PETTERSEN
HENNING PETTERSEN TVEIT

SUPERVISOR
Valeriy Zakamulin

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

Abstract

Whether optimized portfolio strategies have superior performance to the naive diversification or not, has been a heated debate since the publication of DeMiguel, Garlappi, and Uppal (2009). The authors evaluated 14 models which arose as suggestions to solve the issues of estimation errors associated with the mean-variance framework (Markowitz (1952)). They concluded that none of these suggestions consistently outperformed the 1/N benchmark. As a result, several studies followed to claim the opposite. However, more recent critiques have pointed at several issues regarding the testing methodology in these conflicting studies. First, when several strategy performances are tested against some benchmark individually, the results may be affected by data-mining bias. Second, when using Sharpe ratio as the measure of performance, strategies can appear superior due to anomaly effects in the sample of data. We will by our thesis replicate and extend previous research by employing a methodology that considers and deals with the mentioned issues. We will evaluate 9 portfolio strategies against the 1/N portfolio as benchmark. As for testing methodology, we employ White's reality check (WRC) and its extension, the superior predictive ability test (SPA), to avoid data-mining bias. Further, we adopt the Fama/French Carhart's (FFC) 4-factor model to account for potential anomalies in our sources of data. Our findings suggest that when these extensions within methodology are employed, none of the optimized strategies tested are significant for US data. However, when simulating the same strategies in Norwegian returns, our test results report that the three best performing strategies outperform the benchmark significantly.

Acknowledgement

Our thesis is conducted as the final part of our master's program in financial economics at the faculty of business and law at the University of Agder. Through the 4 semesters, we have as students been through extended valuation theory, advanced econometrics, and portfolio theory. Further, we have been practicing the adequate computational and quantitative techniques employed in the theory, besides learning how to use computer-based programs as statistical tools in the introductory coursing to scientific work.

Our deepest and most sincerely gratitude goes to our supervisor Professor Valeriy Zakamulin. Through the master program, he has guided us through both computational theory, computer-based programming and provided us with an introduction to how to conduct scientific work. Further, he has guided us patiently through our master thesis with a personal engagement, which has inspired and motivated us to put as much of our effort as possible into the subject of portfolio management and how to present our adopted knowledge, within the frame of scientific norms.

Further, we want to thank all the other professors we have met on our way for their inspiring enthusiasm and willingness to share their time and knowledge for us to achieve personal growth.

Last but not least, we thank the faculty of business and law and the University of Agder, for in a professional and responsible way, facilitating for and providing us with the achieved experiences, skills and insights.

Table of Contents

List of Tables

1 Introduction

Modern portfolio theory is considered to have its origin from the mean-variance framework of Markowitz (1952), where the optimal portfolio is constructed as the most efficient trade-off between asset's risks and returns. The framework still exists as a fundamental and theoretical guideline for practitioners even today. The theory is working well when the input parameters are known. Though, in practice, it is challenged by potential estimation errors in the predictions of the variance-covariance matrix and future returns. When the inputs deviate from the out-of-sample parameters, the portfolios are characterized by extreme asset weights and poor out-of-sample performances (Merton (1980)).

As a result, a considerable effort in the following research aims to remedy these drawbacks. One approach is to reduce the impact of the potential estimation errors, as by the Bayesian approach (Barry (1974)) and shrinkage estimators (Ledoit and Wolf (2004)). Another approach is to entirely leave out the returns in the computation of asset weights, as for the minimum-variance portfolio (MinVar). However, when the only emphasize on portfolio construction is to minimize risk, another potential problem is the poor diversification of the asset‘s idiosyncratic risks. To emphasis diversification, the maximum diversification portfolio (MDP) minimizes the ratio of asset volatilities to the portfolio volatility, such that the portfolio risk is diversified on several assets (Choueifaty and Coignard (2008)). Further, two other suggested risk-based models, which emphasize diversification, are the equal risk contribution (ERC) and risk parity portfolio (RP). The ERC portfolio diversifies with regard to asset risk. The portfolio is constructed in such a way that each asset contributes equally in terms of risk, to the total risk of the constructed portfolio (Maillard, Roncalli, and Teiletche (2010)). The RP is a simplification of the ERC portfolio, assuming that the correlations between the portfolio's assets are independent. This results in a simplification which has a normalizing effect on the weights, meaning that the total of weights will sum up to 1 (Qian (2011)).

In the wake of the many proposals of mean-variance improvements, DeMiguel, Garlappi, and Uppal (2009) evaluate 14 models, both mean-variance based and risk-based, Bayesian approach and shrinkage estimators. The models are simulated in 7 datasets from Kenneth French's online library of US data, where the 1/N portfolio is used as a benchmark. The conclusion is that none of the approaches are superior to the benchmark when employing three different performance measures, including Sharpe ratio. The results of the study provoked a heated debate towards whether the optimizing framework provided by Markowitz (1952) adds superior value to the naive portfolio or not.

Kritzman, Page, and Turkington (2010) claim that optimized strategies such as the minimum-variance and mean-variance portfolio outperform the 1/N portfolio considering Sharpe ratio in 8 datasets of US data. Tu and Zhou (2011) find that by combining the optimized rules with the rules of the 1/N portfolio, the optimal combinations will outperform the 1/N portfolio in most scenarios. Further, Kirby and Ostdiek (2012) find that two optimized timing strategies outperform the naive portfolio due to the construction of portfolios with low turnover. They claim that high portfolio turnover will erode optimized strategies' performances in cases with high allocation transaction costs.

The commonalities for the mentioned studies is that they were criticized for their use of data and methodology. One commonality is the use of Sharpe ratio as a performance measure. In a recent study by Zakamulin (2017), the author reveals the presence of a low-volatility anomaly in the online library of Kenneth French's US data. This implies that low-risk assets tend to yield greater returns than the high volatility assets. When Sharpe ratio is the employed measure of performance, the performance is based on the simulated returns and their standard deviation. If the assets in the portfolio yield high returns due to anomalies, the Sharpe ratio will not account for these effects. As a result, the Sharpe ratio will give an impression of a greater strategy performance, which in practice should be dedicated to the anomaly effect. Another commonality is the vulnerability to data-mining bias, meaning when several strategy performances are tested individually against some benchmark. The probability of observing significant outperformances and rejecting the null hypothesis despite it is true (type 1 error), will then increase as more individual tests are added.

With these commonalities in mind, Hsu, Han, Wu, and Cao (2018) examine 17 strategies based on the mean-variance framework. The data-mining bias is accounted for by the implementation of White's reality check (WRC) and the superior predictive ability test (SPA). The strategies are tested against the 1/N portfolio as the benchmark in 22 developed country indexes, 30 listed stocks from Dow Jones Industrial Average, and 100 stocks listed in NYSE. The study finds that the 1/N portfolio may perform as well as any of the more sophisticated strategies evaluated in terms of Sharpe ratio and CAPM alpha. In the same context, Yang, Cao, Han, and Wang (2019) study 15376 technical trading rules, and several risk-based strategies are tested against the buy-and-hold benchmark in terms of the 1/N portfolio employing White's reality check and the superior predictive ability test. The study concludes that neither the technical trading rules nor the strategies outperform the buy-and-hold benchmark in terms of mean returns, Sharpe ratios, or certainty-equivalent return.

Our goal is to replicate and extend the previous studies in order to contribute to the ongoing debate, whether there are any optimized strategies that can outperform the benchmark. Our extensions to the previous studies are that we employ the joint testing methodology similar to Hsu et al. (2018) and Yang et al. (2019) to control for the issues of data-mining bias. However, as an extension to Hsu et al. (2018) and Yang et al. (2019), we will employ more factors in our performance measure, to control for data anomalies and consider an additional source of data. The combined tests will be conducted by White's reality check (WRC), where the difference in strategy and benchmark performances will be evaluated against a bootstrapped distribution of outperformances (White (2000)). We also employ the superior predictive ability test (SPA) by Hansen (2005), as an extension of WRC. In the extension, the outperformances are normalized in a parametric test before evaluated against the bootstrapped distribution. Further, we will, in addition to the traditional Sharpe ratio, apply the Fama/French Carhart (FFC) 4-factor model to account for anomalies in the data. The purpose is to avoid having significant results due to factors that are not caused by the strategies. Last, we add a new source of data with returns from the portfolios from the Oslo Stock Exchange, to avoid being influenced by the same data characteristics as the previous studies. In each of the combined tests, we will separate the tests with respect to performance measures (Sharpe ratio and FFC 4-factor alpha) and economic region (US- and Norwegian data). Still, each of the combined tests will include all datasets within each region, along with the tested strategies. Within each economic region's dataset, the strategies are simulated through the historical data, where each strategy allocation is formed on the basis of a 60 month in-sample period. Further, the portfolios are actively managed in accordance with the 9 strategy rules, for each month of the out-of-sample period. Similar to the previous studies, we employ the 1/N portfolio as the benchmark of our study. The 1/N portfolio is recognized for allocating weights equally among the portfolio assets and is therefore not regarding neither asset's expected returns nor variance. In addition to the absence of estimator errors, the 1/N portfolio benefits from its simplicity with regard to both time- and transaction costs. Despite its simplicity, the 1/N portfolio has according to several previous studies, often performed quite well.

For US data, our findings are in accordance with previous studies (DeMiguel et al. (2009), Zakamulin (2017), Hsu et al. (2018), and Yang et al. (2019)) where we do not find that optimized strategies are superior to benchmark. For Norwegian data, however, our results suggest that the 3 best performing strategies outperform benchmark with respect to FFC 4-factor alpha, and we discuss how these results can arise as consequences of market efficiency. The overall conclusion from the summarized results is that the optimized strategies superiority is in general yet to be proved.

The thesis is from this point structured as follows: Section 2 provides a systematic review of relevant literature, from the modern portfolio theory‘s origin, through the development which leads to our specific applied methodology. Section 3 reveals the data sources and from which variables each dataset is sorted by. Section 4 presents the methodology, with a presentation of the strategies, performance measures, bootstrap methodology, and hypothesis tests applied for the different data. Section 5 provides the empirical results obtained from both the White's reality check and the superior predictive ability test. For illustration purposes, we also add results from individual parametric tests, to see which results we would have achieved by traditional hypotheses tests. Section 6 presents the discussion, where we discuss the results in the context of previous literature and our findings during the process. Section 7 summarizes the thesis with a conclusion based on the literature review, applied methodology, and reported results.

2 Literature Review

Modern portfolio theory developed from the publication of the mean-variance framework provided by Markowitz (1952). The theory assumes the desire to maximize future returns while minimizing the portfolio risk. When assets means and variances are known, the theory provides an efficient concave curve of optimal portfolios, known as the efficient frontier. Where to be along this curve depends on the investor's risk aversion. Because the parameters are not known but based on historical returns, the theory has been criticized for not being optimal in practical implementation. Even if there is some stability in the estimated variance, the asset returns have shown to be highly fluctuating and difficult to predict. As a result, the optimized strategies are exposed of estimation errors which results in portfolios with extreme asset weights that perform poorly out-of-sample Merton (1980).

2.1 Optimized Strategies

A broad response within the literature aims to solve the issues of estimation errors. Several of the suggestions aim to reduce the impact of estimation errors by, for instance, the

Bayesian approach (Barry (1974)) and shrinkage estimators (Ledoit and Wolf (2004)). Instead of reducing the impact of estimation errors, another suggested approach is to employ strategies that fully leaves out the returns as input parameters in the computation of weights. The reason is that returns are considered unpredictable due to their fluctuations over time. Variances, on the other hand, are found to be more stable, hence easier to predict.

In an analysis of the risk-based minimum-variance portfolio (MinVar), Clarke, de Silva, and Thorley (2006) find that the variance-covariance risk of the 1000 largest market capitalization stocks is persistent. In addition, they find that the MinVar adds value above the market capitalization-weighted benchmark. Still, without any further consideration of diversification, an obvious drawback with the MinVar is the portfolio's potential concentration around low volatility assets (Maillard et al. (2010)). This results in portfolios which do not fully gain from asset diversification, and the elimination of asset idiosyncratic risks. As Clarke, De Silva, and Thorley (2011) concludes, "with shorting constraint, the optimal portfolio weights decline with increasing market beta, and excludes assets with betas above a certain threshold value."

With the emphasis on portfolio diversification as a criterion for portfolio construction, Choueifaty and Coignard (2008) introduce the maximum diversification portfolio (MDP). The portfolio diversification is measured by the ratio between the weighted average of the assets and portfolio volatilities. The portfolio volatility is computed by including the correlation matrix between the assets, along with the weighted average of asset volatilities. The diversification ratio is then minimized with regard to computing the asset's weights, subject to the constraint that the denominator or portfolio volatility is held constant. If not all of the asset weights' correlations are zero, the diversification ratio will be less than 1, such that the higher the ratio, the higher the diversification.

Maillard et al. (2010) suggests another diversifying strategy known as the equal risk contribution portfolio (ERC). The portfolio‘s weights are constructed such that each portfolio asset, contributes equally to the portfolio's total risk. This implies that if the asset is more volatile than the other, its weight will be proportionally less. The result is a portfolio that is well diversified in terms of risk, where the portfolio weights are greater in assets of low volatility. The conclusion from the evaluation of the ERC portfolio against the 1/N benchmark is; when Sharpe ratio is employed as the measure of performance, the benchmark is outperformed. However, the evaluation also finds that the ERC portfolio can be outperformed by the MinVar due to the MinVar‘s low portfolio volatility. In other words, the ERC portfolio does not always outperform the naive- and MinVar portfolios. Still, due to their high level of diversification, these constructions provide more stable portfolios, less vulnerable to volatile market conditions. The risk parity portfolio (RP) is a simplification of the ERC portfolio, and the similarities between the two are obvious. Still, the difference is that the RP portfolio assumes that the returns between the portfolio's assets are independent. This results in a simplification which has a normalizing effect on the weights, meaning that the total of weights will sum up to 1 (Qian (2011)).

As a consequence of the issues related to estimation errors in the framework of Markowitz's theory, tedious efforts are made to enable practical implementations. DeMiguel et al. (2009) evaluate 14 of these various methods and strategies against the 1/N portfolio as the benchmark. The strategies are based on the rules of the mean-variance framework, such as the classical mean-variance portfolio, which optimizes returns per given level of risk: Further, the authors evaluate suggested extensions of the framework, which aim to solve the issues of estimation errors in the Markowitz framework. The MinVar, which leaves out the estimation of returns, and the various methods which aims to reduce the impact of the estimation errors, such as the Bayesian approach and moment restrictions. Their conclusion is that none of these approaches outperforms the naive portfolio significantly. Due to these controversial findings, a heating debate raises several studies which claim the superiority of the mean-variance based portfolios.

Kritzman et al. (2010) apply Sharpe ratio as a performance measure and conduct a study of 13 datasets and over 50.000 constructed optimized portfolios, which they compare with the market portfolio and the 1/N portfolio. They argue that poor estimation models, due to short historical samples of 5 years, explains why previous research did not argue in favor of mean-variance portfolios. The conclusion is that optimized portfolios outperformed the passive benchmark when the estimation of parameters was based on a longer rolling window horizon. Tu and Zhou (2011) find that by combining the optimized rules with the rules of the 1/N portfolio, the optimal combinations will outperform the 1/N portfolio in most scenarios. Further, that the contrary conclusions from previous studies are due to estimation errors in the applied methods. Kirby and Ostdiek (2012) find that two optimized timing strategies, volatility timing portfolio, and the risk-to-reward portfolio outperforms the naive portfolio due to the construction of portfolios with low turnover. The timing strategies were simulated through 4 different sets of data. Further, they claim that high portfolio turnover will erode optimized strategies' performances in cases with high allocation transaction costs. To deal with this, they suggested a strategy that ignores the asset correlations and consider only risk and return in the computation of asset weights. To minimize estimation risk, the strategy is added a tuning parameter, which decides how the allocation of weights shall respond to asset risks.

Despite several studies claiming defense of Markowitz's portfolio theory, Zakamulin (2017) reveals two major commonalities in these previous studies. First, with regard to the use of French's online library of US data. Second, in the choice of Sharpe ratio as the applied performance measure. Zakamulin (2017) demonstrates the existence of low volatility anomaly in the datasets of US data and points out how Sharpe ratio does not account for these effects. In contrast to the rewarded-risk intuitivism within the mean-variance framework, the low-volatility anomaly rewards the less risky assets. Since Sharpe ratio does not account for these effects, there will potentially be strategies which tilts against low volatility assets, that outperforms the 1/N benchmark due to the anomalies, and not due to the strategies themselves. As we know, the 1/N portfolio is distributing weights equally among portfolio assets without regarding either volatilities or returns. An additional commonality in the for mentioned previous studies is the probability of receiving significant outperformances due to data-mining bias. This phenomenon occurs when several strategies are evaluated individually against some benchmark, and not in combined tests with the other strategies evaluated jointly. When testing several strategies individually, the probability of getting significant results due to type 1 error rather than the strategies actual outperformance increases with the number of strategies tested (White (2000)).

A study that controls for the data-mining bias is Hsu et al. (2018). The examination employs the White's reality check (White (2000)) combined with Hansen's superior predictive ability test (Hansen (2005)) in the evaluation of 16 basic strategies, 126 learning strategies, and approximately 2000 extended strategies. First, the study examines the strategies in conventional tests and finds indications of superiority against the 1/N benchmark in some of the strategies. However, when they apply the WRC and SPA test, which controls for data-mining bias, the results are in the opposite direction, and in accordance with DeMiguel et al. (2009). The study does, however, only use Sharpe ratio and the CAPM alpha as performance measures, which on Kenneth French's online library of US data will be unable to account for the anomaly effects as pointed out earlier.

In a similar study with regard to the use of joint tests, Yang et al. (2019) evaluate 15376 technical trading rules and several strategies suggested in the literature. From the mean-variance framework, the authors employ the mean-variance portfolio (MVP) along with the MinVar. From the risk-based universe, they add the maximum diversification portfolio (MDP) (Choueifaty and Coignard (2008)), the ERC portfolio (Maillard et al. (2010)), volatility timing portfolio (VolTim) and reward-to-risk timing portfolio (RRTim) (Kirby and Ostdiek (2012)). The strategies are tested against the buy- and hold benchmark in terms of the 1/N portfolio, employing White's reality check and the superior predictive ability test. The study concludes that neither the technical trading rules nor the strategies outperform the buy-and-hold benchmark. The performance measures used are the factor-free mean returns, Sharpe ratio, and certainty-equivalent return.

2.2 Asset and Portfolio Pricing

Sharpe (1964), Treynor (1962), Lintner (1965), and Mossin (1966) developed the CAPM, which connects the asset's expected returns to their sensitivity to the general market. Black, Jensen, and Scholes (1972) further develop the model. The motivation behind the modification is empirical evidence for asset returns not being strictly proportional to the asset's betas. The empirical testing reveals a different pattern in terms of beta dependency. The phenomenon is later confirmed empirically by Haugen and Heins (1972) and is today referred to as the low beta anomaly. Fama and French (1993) extend these studies and collect average returns on stocks from NYSE, AMEX, and Nasdaq. Their examination is a cross-sectional regression study with beta, size, E/P-ratio, leverage and book-to-market as potential risk factors. In the conclusion, they do not find evidence for supporting the earlier assumption towards a beta driven asset return. Instead, they claim that risk factors as size and book-to-market equity are explanations of return anomalies. The outcome is the introduction of a multifactor regression model, the three-factor model. The model adds factors for MKTMRF which is the market excess return to the risk-free rate, SMB which is small minus big (with respect to firm size based on market capitalization), and HML which is High minus low (with respect to book-to-market ratios). Carhart (1997) extends the Fama/French 3-factor model with one extra factor, the momentum factor. He finds that the extension of one extra factor explains the US market better than the original 3-factor model. The momentum factor is based on the findings of Jegadeesh and Titman (1993), where they capture the one-year momentum anomaly.

3 Data

The data in our thesis comes from two different sources.

First, from Kenneth R. French's online data library of US data[1], from where we have chosen 15 arbitrary datasets with different starting dates, but where none of them starts later than July 1963. The data consist of monthly returns, where each dataset is constructed by 10 assets, sorted by different variables, from the lowest to the highest 10th decile in accordance with the respective variables. The return data are collected from the US stock indexes; Nasdaq, Amex, and NYSE.

Second, from Bernt Arne Ødegaard's online data library of Norwegian data[2], from where we have chosen 4 datasets with starting dates January 1980 and 1981. However, since we need the corresponding risk-free rates, we set the starting dates to July 1981. As for the US data, the Norwegian datasets consist of 10 assets sorted by different variables shown in the list of strategies. The Norwegian data are monthly returns, as for the US data, only now collected from Oslo Stock Exchange.

Table 1 shows the list of all datasets, where all datasets consist of value-weighted portfolios, due to findings and recommendations from Plyakha, Uppal, and Vilkov (2014).

3.1 Factors

For the tests using 4-factor alpha as performance measures, we have downloaded the monthly risk-free rates (RF) along with the factors for market excess returns (MKTMRF), small minus big firms (SMB), high minus low book-to-market ratios (HML) and momentum factor (MOM) which will be discussed further in Section 4. The factors for US data are collected from the French database as for the return datasets. For Norwegian data, the factors are collected from the same source as for Norwegian return data.

[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`

Market	#	Dataset	N	Start	End	Abbreviation
USA	1	Size	10	1926-07	2019-12	Size
	2	Book-to-Market	10	1926-07	2019-12	B/M
	3	Operating Profitability	10	1963-07	2019-12	OP
	4	Investment	10	1963-07	2019-12	Inv
	5	Momentum	10	1927-01	2019-12	Mom
	6	Short-Term Reversal	10	1926-02	2019-12	STR
	7	Long-Term Reversal	10	1931-01	2019-12	LTR
	8	Accruals	10	1963-07	2019-12	ACC
	9	Market Beta	10	1963-07	2019-12	Mkt B
	10	Net Share Issues	10	1963-07	2019-12	NSI
	11	Residual Variance	10	1963-07	2019-12	ResVar
	12	Variance	10	1963-07	2019-12	Var
	13	Earnings/Price	10	1951-07	2019-12	E/P
	14	Cashflow/Price	10	1951-07	2019-12	CF/P
	15	Dividend Yield	10	1927-07	2019-12	DivY
Norway	16	Book-to-Market	10	1981-01	2018-12	B/M
	17	Size	10	1980-01	2018-12	Size
	18	Momentum	10	1980-01	2018-12	Mom
	19	Spread	10	1981-01	2018-12	Spread

Table 1: This table lists the analyzed datasets, from which characteristics they are sorted, the number of assets in each dataset and their time periods. In the last column we have listed the abbreviations used in our tables.

4 Methodology

4.1 Portfolio Strategies

In this section, we discuss the different portfolio strategies listed in Table 2. We assume there are N risky assets, where w_i is the weights of each asset, and i denotes the i_{th} asset. r_f is the notation for the risk-free asset, $\boldsymbol{w}$ is the weight vector, $\mathbf{w}'$ is the transposed weight vector, $\boldsymbol{\Sigma}$ is the variance-covariance matrix, and $\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}$ is the notation for the portfolio's variance. $\boldsymbol{\mu}$ is a vector of returns, and $\mathbf{1}$ and $\mathbf{0}$ are vectors of respectively ones and zeros.

#	Name	Abbreviation
1	Naive Diversification	Naive
2	Mean-Variance	MVP
3	Minimum-Variance	MinVar
4	Maximum Diversification	MDP
5	Equal Risk Contribution	ERC
6	Risk Parity	RP
7	Volatility Timing	VolTim
8	Reward-to-Risk Timing	RRTim
9	Maximum Decorrelation	MaxDec
10	Risk-Efficient	RiskEff

Table 2: This table lists the various portfolio strategies we consider. In the last column we have the abbreviation we use to refer to the strategies in our tables.

4.1.1 Naive Diversification

The naive portfolio, also known as the 1/N portfolio, diversifies an equal percentage share into each of the portfolio‘s assets. Consequently, the composition of the 1/N portfolio does not depend on asset returns and standard deviations in the computation of asset weights. Despite its simplicity, the naive diversification has often been found to perform at least as good as optimized strategies in several of the previous and relevant studies, e.g., DeMiguel et al. (2009). The combination of these features results in a portfolio that both yield high performance and require a minimum of resources, as time spent and transaction costs. As a result, the 1/N portfolio will, as in many of the previous studies, constitute our benchmark portfolio. The 1/N portfolio's weights are computed as follows:

$$w_i = 1/N \tag{1}$$

4.1.2 Mean-Variance

The mean-variance portfolio (MVP) is rooted in the portfolio theory of Markowitz (1952), which is regarded as the origin of modern portfolio theory. The principal idea behind Markowitz's introduction was to maximize portfolio mean returns for each level of portfolio risk. For each portfolio's composition of weights, the optimal portfolios are represented by the efficient frontier. The curve starts in the point of lowest volatility possible and stretches outwards as the volatility increases. The optimal point on this curve depends on whether the investor considers risk-free assets as available or not. If there is a risk-free asset, the point on the efficient frontier is determined by the construction of a tangency line between the risk-free asset‘s level of return and the tangency point to the efficient frontier. The tangency point (market portfolio) is then the optimal portfolio. The portfolio‘s weights are computed as:

$$\min_{\mathbf{w}} \frac{1}{2}\, \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \text{ s.t. } \mathbf{w}'(\boldsymbol{\mu} - \mathbf{1}r_f) = \mu^\star - r_f,\ \mathbf{w} \geq 0, \tag{2}$$

where $\mu^\star$ is desired return.

An important assumption of the MVP is that the portfolio‘s mean return is greater than the risk-free rate. If not, the CAL will be negative. In practical situations, this assumption is not always met. To avoid this exposure, we choose an approach where the efficient frontier is determined by the investor’s risk aversion. The risk aversion is used as a coefficient, which we have pre-set to 1000. In the case of no borrowing restrictions, the coefficient can be any arbitrary number, without affecting the portfolio weights. In our thesis we assume borrowing restrictions and set the coefficient to be a high number, to ensure getting the same results as for the MVP portfolio. The risk aversion coefficient is then applied in the formula where the investor’s utility is maximized. By Zakamulin (2017) disclosure of low-beta anomalies in US data, we saw that risk-minimizing strategies outperformed the 1/N benchmark in terms of Sharpe ratio. The consideration of the MVP is interesting because we are now able to test the strategy against the benchmark in the light of low volatility anomaly using the FFC 4-factor model. With the adoption of a risk aversion coefficient, the weights are now computed as:

$$\max_{\mathbf{w}} U(\mathbf{w}) = (\boldsymbol{\mu} - \mathbf{1}r_f)'\mathbf{w} + r_f - A\frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \text{ s.t. } \mathbf{w} \geq 0, \tag{3}$$

where U is utility, and A is the risk aversion coefficient.

4.1.3 Minimum-Variance

The minimum-variance portfolio (MinVar) is as the name reveals, the portfolio on the efficient frontier with the lowest volatility, situated on the very left tip of the curve on the variance scale. The strategy arose from the issues of potential estimation errors in the predictions of future returns. As a result, the MinVar portfolio is only considering variance as input parameter, which is considered a more stable parameter to predict. In common with the MVP, the MinVar also base asset allocation on minimization of portfolio risk, but without regards to returns. The weights of the minimum-variance portfolio are computed as:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\mathbf{\Sigma}\mathbf{w} \text{ s.t. } \mathbf{w}'\mathbf{1} = 1 \text{ and } \mathbf{w} \geq 0. \tag{4}$$

Clarke et al. (2006) argue that the purpose of asset diversification is to minimize risk per given return. Since the returns cannot be properly estimated, a solution is to use the variance-covariance matrix of assets. The asset weights will then be computed such that the portfolio variance is as low as possible. In addition, they claim that the MinVar portfolio has excess returns that tangent or exceeds the historical market. However, a problem that often occurs with the MinVar is that the portfolio concentrates too much weights around the low volatility assets, resulting in a poorly diversified portfolio. The issue of poor diversification has initiated extensions where the diversification aspect has been in focus. We regard the MinVar portfolio as an essential strategy to include in our hypothesis test for several reasons. First, this is the strategy with the lowest perceived risk. Second, contrary to the benchmark, the MinVar portfolio potentially selects few assets in the portfolio. Third, the MinVar portfolio has been found to outperform the 1/N portfolio in previous studies when Sharpe ratio is applied as a performance measure, e.g., Kritzman et al. (2010). Now the strategy will be evaluated when using FFC 4-factor alpha.

4.1.4 Maximum Diversification

The maximum diversification portfolio (MDP) seeks to solve the MVP‘s drawback with poor diversification by maximizing the portfolio’s diversification. The method ignores expected returns in the computation of weights and focuses only on the standard deviation and correlations between the portfolio assets (Choueifaty and Coignard (2008)), where the weights are computed as:

$$\max_{\mathbf{w}} \frac{\mathbf{w}'\boldsymbol{\sigma}}{\sqrt{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}} \text{ s.t. } \mathbf{w}'\mathbf{1} = 1 \text{ and } \mathbf{w} \geq 0, \tag{5}$$

where $\boldsymbol{\sigma}$ is vector of standard deviations.

The numerator is the transposed vector of weights multiplied by each asset‘s standard deviation, where the result is a vector of the weighted averages of risk from each asset. The numerator is then divided on the portfolio risk, where the result is a ratio between the weighted average of risk and the total portfolio risk. The ratio is then maximized with respect to weights, to find the vector of weights that gives maximum diversification between the assets, given the implemented constraints that the denominator is held constant.

4.1.5 Equal Risk Contribution

The philosophy behind the equal risk contribution portfolio (ERC) arose from the traditional 60/40 portfolio of respectively stocks and bonds, where, even though it may seem like a balanced construction, the risk from the 60% share, contributed with about 90-95% of the total risk. In other words, not balanced in terms of distributed risk (Qian (2011)). The logic behind this strategy’s way of diversifying risk is to construct a portfolio where each asset contributes equally in terms of volatility to the portfolio’s total risk. If one asset is more volatile than the other, the asset will have a proportionally smaller degree of weight. The ERC portfolio is a mean-free portfolio, meaning that it does not depend on asset‘s mean returns in the computation of portfolio weights. By leaving out the mean returns as input parameter, the estimation process gets more robust and reliable in terms of future predictability. The motivation behind including the ERC portfolio in our thesis is the increased attention it got in the wake of the financial crisis, and that Maillard et al. (2010) claims its superiority against the 1/N benchmark portfolio. There is, however, no closed-form solution of the ERC portfolio, but one solution that requires Sequential quadratic programming is the following:

$$\min_{\mathbf{w}} \sum_{i=1}^{N}\sum_{j=1}^{N}(w_i(\mathbf{\Sigma w})_i - w_j(\mathbf{\Sigma w})_j)^2 \text{ s.t. } \mathbf{w}'\mathbf{1} = 1 \text{ and } \mathbf{w} \geq 0, \tag{6}$$

where $\mathbf{\Sigma w}_i$ is the i_{th} weight of the vector $\mathbf{\Sigma w}$.

4.1.6 Risk Parity

The difference between the risk parity portfolio (RP) and the ERC portfolio is that the RP assumes that the correlations between the assets are zero. This assumption simplifies the computation of weights to the following expression (Chaves, Hsu, Li, and Shakernia (2012)):

$$w_i = \frac{\frac{1}{\sigma_i}}{\sum_{i=1}^{N} \frac{i}{\sigma_i}}, \tag{7}$$

where σ_i is the standard deviation of asset i.

We see that the weights are determined by the inverse of each asset‘s standard deviation, divided on the sum of all asset‘s standard deviation inverted. This has a normalizing effect on the weights, meaning that the assets in the strategy sum up 1, where each asset weight is a share of that sum. Except from that, RP is based on the same principles as the ERC portfolio, and benefit from the same attributes with regard to using standard deviation as input and that risk is well diversified among portfolios’ assets, rather than just spreading the weight equally. This results in portfolios where the riskier assets are reduced in weights proportions, while the less risky assets are accordingly increased. Both ERC portfolio and RP are relevant as tested strategies against the benchmark due to their commonalities with the 1/N portfolio with regard to equality in diversifications. The difference is that while the 1/N portfolio diversifies by equal weights, the ERC portfolio and RP diversify by equal risk contributions, where low volatility assets get proportionally larger weights. Due to the low volatility anomalies in US data, this would indicate better performance by the strategies which allocate more into the low volatility assets when using a performance measure (Sharpe ratio), which does not consider the anomaly effects.

4.1.7 Volatility Timing

The motivation from including volatility timing into our study is that Fleming, Kirby, and Ostdiek (2001) & Fleming, Kirby, and Ostdiek (2003) claim that it outperforms the mean-variance strategy. Moreover, Kirby and Ostdiek (2012) found that it also outperforms the naive diversification. The strategy does not depend on returns, but on the assets individual variance, which is tuned into the parameter η in the following expression:

$$w_i = \frac{(\frac{1}{\sigma_i^2})^\eta}{\sum_{i=1}^{N}(\frac{1}{\sigma_i^2})^\eta}, \tag{8}$$

where σ_i^2 is the variance to asset i, and η is a parameter adjusting for how aggressively an investor adjusts weights in response to volatility changes. We have set η to 4 in our study, as Kirby and Ostdiek (2012) found that to be the best number.

4.1.8 Reward-to-Risk Timing

The reward-to-risk timing is closely related to volatility timing. In the reward-to-risk timing portfolio, we are considering excess returns in the composition of weights. The benefit of implementing returns is the possibility of improving portfolio performance. A weakness, however, is the risk of estimation errors due to the fact that returns are estimated with less precision than variance. Another precaution to take is if the asset's mean returns lie beyond the risk-free rate. We will then end up with negative excess returns, which again will result in extreme asset weights. To prevent this, we maximize between the excess return and zero to get a vector of excess returns where zero is the potentially lowest value, such that we do not risk getting negative excess weights, due to the short sale constraint ((Kirby and Ostdiek,2012)). The weights are given by:

$$w_i = \frac{(\frac{\mu^+}{\sigma_i^2})^\eta}{\sum_{i=1}^N (\frac{\mu^+}{\sigma_i^2})^\eta}, \tag{9}$$

where μ^+ is max $(\mu_i - r_f, 0)$, μ_i is return of asset i, and we again set η to 4.

4.1.9 Maximum Decorrelation

The maximum decorrelation portfolio's formula has close similarities with the minimum-variance portfolio and maximum diversification portfolio. The main difference is that the maximum decorrelation, as the name reveals, includes the assets correlation matrix in the computation of asset weight. In contrast minimum-variance and maximum diversification use the covariance matrix. Maximum decorrelation ignores the assets volatilities and mean returns. Hence, the assumption behind this optimization is that the assets returns and risks are the same, or not expedient to predict due to high probability of estimation errors. By avoiding having too many factors to predict based on historical data, the assumption is that this will reduce the estimation error when computing the optimal allocation strategy (Christoffersen, Errunza, Jacobs, and Langlois (2012))). The weights are given by the following expression:

$$\min_{\mathbf{w}} = \mathbf{w}'\mathbf{C}\mathbf{w} \text{ s.t. } \mathbf{w}'\mathbf{1} = 1 \text{ and } \mathbf{w} \geq 0, \tag{10}$$

where $\mathbf{C}$ is the correlation matrix.

4.1.10 Risk-Efficient

The risk-efficient portfolio is sorting the assets by their semi-deviation as opposed to standard deviation, meaning sorting the assets after their worst-case performances. In other words, the assets are sorted by their downside risks. The strategy then computes decile portfolios (divided into ten equal parts), and from that, find the median semi-deviation of assets in each decile portfolio. The strategy takes base in the MVP portfolio with respect to maximizing the risk/reward ratio, but with a different interpretation of risk, which is associated with the MVP portfolio (Amenc, Goltz, Martellini, and Retkowsky (2011)). The risk efficient portfolio is computed as:

$$\mathbf{w} = \max_{\mathbf{w}} \frac{\mathbf{w}'\mathbf{J}\varepsilon}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} \text{ s.t. } \mathbf{w}'\mathbf{1} = 1 \text{ and } \mathbf{w} \geq 0, \tag{11}$$

where $\mathbf{J}$ is a $(\mathbf{N} \times \mathbf{10})$ matrix of zeros whose $(i, j)_{th}$ element $= 1$, if the semi-deviation of stock i belongs to decile $j, \varepsilon = (\varepsilon_1, ..., \varepsilon_{10})$, where ε is a vector of the error term.

4.2 Performance Measures

4.2.1 Sharpe Ratio

As an overall aim by investing in assets is to achieve an as high return as possible for any given risk, the Sharpe ratio (SR) is a widely preferred measure of assets performances. The Sharpe ratio provides the slope of the capital allocation line, which is the linear relation between our strategies' excess returns and the excess returns' standard deviation. The Sharpe ratio is given by:

$$SR = \frac{\mu_p - r_f}{\sigma_p}, \tag{12}$$

where μ_p and σ_p are respectively mean return, and standard deviation of portfolio p, while r_f is the return of the risk-free asset.

We will apply the Sharpe ratio as a strategy performance measure to be able to compare our results with previous studies. Despite its simplicity and applicable advantages, the Sharpe ratio has several shortcomings. One is that the standard deviation in the denominator assumes normally distributed returns. Whenever volatility is positively skewed, the Sharpe ratio misses the fact that the risk first and foremost is associated with positive returns. Further, the Sharpe ratio does not reflect anomalies in the data. If a strategy significantly outperforms the benchmark, the outperformance can be due to these anomaly effects, rather than the strategies themselves. Still, when using the Sharpe ratios in the hypothesis tests as supplements to the FFC 4-factor alpha, we achieve inferences drawn on a broader basis, where the performance measures partly complete each other.

4.2.2 Fama/French Carhart 4-factor model

The Fama/French Carhart 4-factor model arose through a development where the beta in CAPM was empirically tested and found not to be strictly proportional to the asset's returns (Haugen and Heins (1972)). This implies that returns have additional explanations than just their sensitivities to the market return. The deviation between the asset's and market returns are in the CAPM framework designated alpha. As a result, Fama and French (1993) developed the Fama/French 3-factor model, which included two additional explanatory factors. In addition to the asset's sensitivity to the market's systematic risk, the authors added factors for; "small firms' returns - big firms' returns" (SMB) and "high book-to-market equity firms returns - low book-to-market equity firms returns" (HML). The rationale for these inclusions was the disclosure of how these two factors, besides MKTMRF, also captured some significant fractures of the variation of stock returns. Later, Carhart (1997) extends the 3-factor model where an additional factor for momentum is added, to potentially better explain anomalies in US data. The momentum (MOM) is a factor that explains an asset's tendency to keep increasing if the returns increased in the last period, and to keep decreasing if the price decreased in the previous period. The FFC 4-factor alpha is given as the intercept in the following expression:

$$r_{p,t} = \alpha_p + \beta_1 \mathrm{MKTMRF}_t + \beta_2 \mathrm{SMB}_t + \beta_3 \mathrm{HML}_t + \beta_4 \mathrm{MOM}_t + \epsilon_{p,t}, \tag{13}$$

where $r_{p,t}$ is the excess return to the portfolio p at period t, α_p denotes the pricing error to the portfolio, β_i is the exposure measure to the different factors at period t, and $\epsilon_{p,t}$ is the error term to the portfolio at time t which is not explained by the model.

The alpha has now become the assets or portfolios' excess returns to the market portfolio when these firm characteristics are controlled for. In our thesis, this results in multivariate regressions, where the strategies' alphas are computed before the strategies' alphas are paired against the benchmark's alpha. The 4-factor model is applied in our hypothesis test to control for the SMB-, HML- and MOM-anomalies, with concerns to the findings of return anomalies in the US data (Zakamulin (2017)).

4.3 Out-of-sample-testing

When constructing a portfolio, the composition is based on the portfolio strategy used, and the parameters' expected values. For our purposes, these parameters are based on historical data, specifically, from what we define as the optimized or in-sample period (1 to t) in the out-of-sample-test of historical returns. An out-of-sample-test is the simulation of returns when we look at how a strategy would have performed in the past, pretending not to know about its future. When compositions are set based on in-sample data, the period length will stay constant as a rolling window, moving one month ahead at the time. At time t, the out-of-sample period starts. At $t+1$, the moving window has moved one month ahead and will re-allocate its portfolios weights in accordance with the designated strategy and the new windows parameters. Then the same will occur in time $t+2$ and so on. In our thesis, the length of our in-sample period is set to 60 months, to stay within the frame of similar research studies. The in-sample period is set from July 1963 to July 1968 for US datasets, and from June 1981 to June 1986 for the

Norwegian datasets. The out-of-sample periods will start from the respective datasets in-sample periods ends, to the end of 2019 for US data, and end of 2018 for Norwegian data. From the strategy simulation, we will, in the end, be able to extract the strategies' and benchmark's mean returns, along with their respective standard deviations. The parameters will then suit as inputs for the computed performance measures.

4.4 Hypothesis Tests

Many of the test statistics used in inferences about sample realizations are based on the fulfillment of certain assumptions towards the distribution's characteristics. For asset return data, the literature has shown that return data often do not fulfill assumptions towards normality, stationarity, and independency. These tendencies make traditional parametric hypothesis tests less robust. As a response, many researchers have tried to come up with statistical solutions to solve these challenges when predicting the future based on historical returns. One solution is the construction of distributions based on the resampling process of asset returns many times, in a methodology known as bootstrap resampling.

Another more critical challenge in traditional testing is the phenomenon known as the data-mining bias or data snooping bias. The term data-mining comes from the methodology where the goal is to reveal patterns in great amounts of data. In our context, data-mining refers to the effort in finding the best performing strategy among several other strategies in a large amount of return data.

In traditional testing, the strategies are tested individually against the benchmark in a z-test for Sharpe and a t-test for alpha. The z-test is given as:

$$z = \frac{SR_k - SR_{BM}}{\sqrt{\frac{1}{T}[2(1-\rho+\frac{1}{2}(SR_k^2+SR_{BM}^2-2SR_kSR_{BM}\rho^2)]}}, \tag{14}$$

where SR_k is the Sharpe ratio for the active strategy, SR_{BM} is Sharpe ratio for the benchmark, T is the number of monthly observations, and ρ is the correlation between the returns of the active strategy and the benchmark. Further, the t-test is computed as:

$$t = \frac{\alpha_k - \alpha_{BM}}{\sqrt{\frac{1}{T}\sigma_k^2 - 2\rho\sigma_k\sigma_{BM} + \sigma_{BM}^2}}, \tag{15}$$

where σ_k is the standard deviation for the active strategy, σ_{BM} is the standard deviation for the benchmark, α_k is the alpha value for the active strategy, and α_{BM} is the alpha value for the benchmark. The p-values from the individual parametric tests are then given as:

$$p_s = \text{prob}(z > z_{1-\alpha}),$$

where p_s is the p-value of a single test, z is the realization from the individual performance test, $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution and α is the significance level. If $\alpha = 5\%$, the probability of rejecting the null hypothesis even when it is true, is 5% (type 1 error). However, when we are evaluating several strategies in several hypothesis tests, this probability of type 1 error increases dramatically. As we can see from the following formula, the p-values are now given as:

$$p_N = 1 - \text{prob}(z_1 < z_{1-\alpha}; z_2 < z_{1-\alpha}; ...; z_N < z_{1-\alpha}) = (1 - p_S)^N,$$

where p_N is the p-value of the multiple tests, and z_i is the realization from the individual performance test. In our case, with 9 evaluated strategies, the probability of rejecting the null hypothesis due to type 1 error would have been approximately 37% in terms of traditionally testing as demonstrated below:

$$1 - (1 - 0.05)^9 = 0.3698 \approx 37\%.$$

The phenomenon of data-mining bias has long been acknowledged in the literature. A suggested solution is by White (2000), where White's reality check (WRC), or just the reality check (RC) is presented. Due to this, we will first test the US data with regard to Sharpe ratios applying White's reality check. After the tests of Sharpe ratios in US data, we will continue with the same test on the Fama/French Carhart 4-factor model. For the Norwegian data, we will follow the same steps as for US data. As an extension of the WRC, we will further apply the suggested SPA test by Hansen (2005) for the same performance measures, in the US- and Norwegian data. For both combined hypothesis tests, the strategies excess returns will be bootstrapped to achieve their distributions, without relying on assumptions towards their true values. For our purposes, we will use a block bootstrap method which resamples returns in blocks of certain lengths to deal with dependency between the returns due to time-series autocorrelation (Politis and Romano (1994)). The mean block length (b_{len}) is determined by taking the mean of the vector of the benchmark strategy's optimal block lengths. The mean is used to compute a probability parameter ρ, where:

$$\rho = \frac{1}{b_{len}}. \tag{16}$$

The probability of drawing blocks of k observations is then:

$$P(k) = \rho(1-\rho)^{k-1}. \tag{17}$$

We then start at a random point in the strategies' excess returns from time t_1, t_2, ..., t_n, and for each t_n, we flip a biased coin with probability ρ. In the event of success, the block is continued; otherwise, a new random block is started. From the geometrically distributed blocks, we bootstrap distributions with stationarity in the data (Masters (2018)). In practice, when applying joint hypothesis tests with respect to Sharpe ratio and 4-factor alpha in each of our datasets, we perform the hypothesis test combined. By combined, we, in addition to combining the different strategies within each dataset, also combine the strategies' excess returns from all datasets. One of the advantages of testing several datasets and strategies combined is that it opens for a significantly larger sample size when seeking to find universality in, for our purposes, strategy outperformance.

4.5 White's Reality Check

4.5.1 Sharpe Ratio

As Halbert White (2000) stated, several studies have highlighted their concerns towards joint test inferences; however, few attempts have been published in order to derive an applicable method, which was the case and focus of White's study.

The starting point for our hypothesis tests is to see if the best strategy can outperform benchmark. In addition, we will also see if the second, third, and fourth best strategies can outperform the benchmark. We will present the methodology for the best strategy, but the methods for the other best strategies are the same. For each strategy, the mean excess returns and their standard deviation are computed to construct a vector of Sharpe ratios for each strategy. The vector is then subtracted the benchmark's Sharpe ratio, which results in a $1xk$ vector of outperformance measures:

$$f_k = SR_k - SR_{BM}, \tag{18}$$

where f_k is the strategies outperformance to the benchmark, SR is Sharpe ratio, k refers to the k_{th} strategy, and BM is the benchmark.

The best observed outperformance is then selected as f_{max}:

$$f_{max} = \max_{k=1,\ldots,n} f_k, \tag{19}$$

where n is the number of strategies

The null hypothesis is that the best observed performance does not outperform the benchmark:

$$H_0 : f_{max} \leq 0. \tag{20}$$

Due to the issues with data-mining bias, as derived earlier, the following steps will follow White's reality check (White (2000)), which adopts the stationary bootstrap for dependent data. By resampling the strategies' observed excess returns, we are now able to compute distributions of Sharpe ratios for each strategy $SR_{j^\star,k}$, and for the benchmark $SR_{j^\star,BM}$. The block length is determined by the method described in Politis and White (2004), with geometrically distributed block length as derived earlier in the methodology section.

The bootstrap resampling will, in our test, be repeated 1000 times, where for each resample, we compute the resampled strategy outperformances against the benchmark. The number of times we resample is based on repeated attempts with different numbers of resamples, from which we observed that when resampling above 1000 times, our p-values stabilized, meaning, they stopped being influenced by a further increase. In addition, choosing $J = 1000$ as the number of resampling times also has support in the relevant literature.

Hence, from the 1000 times resampled returns for each strategy and the benchmark, their Sharpe ratios are computed. The strategies' ratios are then subtracted the benchmark's ratio, and we are left with strategy distributions of outperformance measures over the benchmark:

$$f^\star_{j,k} = SR^\star_{j,k} - SR^\star_{j,BM}, \tag{21}$$

where $f^\star_{j,k}$ is the strategy distributions of outperformances against the benchmark, and j is the j_{th} resample.

The next step is now to select the best outperforming strategy. This leaves us with a vector of outperformances against the benchmark. To comply with the null hypothesis stating that the best strategy does not outperform the benchmark, each of these outperformances is now subtracted the strategies' observed outperformance f_k.

$$f^{\star}_{j,max} = \max_{k=1,\dots,n} (f^{\star}_{j,k} - f_k). \tag{22}$$

Last, we calculate the p-value by counting how many times $f^{\star}_{j,max}$ is higher than the observed f_{max}, where this number is divided on the number of resamples J:

$$p_{value} = \sum_{j=1}^{J} \frac{\mathbf{1}_{f^{\star}_{j,max} > f_{max}}}{J}, \tag{23}$$

where J is 1000 and **1** denotes the indicator function that takes the value of one of the conditions are fulfilled.

If the observed best outperformance f_{max} beats the bootstrapped best outperformance $f^{\star}_{j,max}$ in more than 95% of the times, the p-value is $1 - 0.95 = 5\%$, and we can reject the null hypothesis stating that the best strategy does not outperform the benchmark.

4.5.2 4-Factor Alpha in WRC

For the 4-factor alpha, we will compare the strategies' alpha intercepts against the benchmark portfolio's alpha intercept. Multiple regression will compute the 4-factor alphas for the strategies and benchmark, where we by adding the time-series data for portfolio's excess returns, market excess return, small minus big, high minus low and momentum, will receive the portfolio's alpha intercepts, as explanations to portfolio excess returns, other than the four explanatory variables.

The respective intercepts or alphas are then subtracted the passive benchmark's alpha, where the result is a vector of strategies' outperforming alphas which later is maximized:

$$f_k = \alpha_k - \alpha_{BM}, \tag{24}$$

where f is the outperformance, k refers to the k_{th} strategy, and BM is the benchmark.

The remaining procedure for 4-factor alphas when using WRC as the hypothesis test follows the same steps as for Sharpe ratio.

4.6 Hansen's SPA test

Although White's reality check deals with the issues regarding data-mining bias, Hansen (2005) proposes an extended version known as the Superior predictive ability (SPA) test. The SPA test suggests two improvements of White's reality check. One is the normalization of the computed observed outperformances for each of the strategies. The outperformances are divided on their standard deviation. That is, instead of using the formula in WRC as $f_k = SR_k - SR_{BM}$, we use the following formula in the SPA test:

$$z_k = \frac{f_k}{\hat{\sigma}_{f_k}}, \tag{25}$$

where z_k is the test statistic, and $\hat{\sigma}$ is the standard deviation of f_k.

This results in strategy outperformances with similar units in the standard deviations. Before the bootstrap is ran, the observed outperformances are computed by a z-test for Sharpe ratios and t-test for alphas, where both provide vectors of strategy z-values. After selecting the best observed strategy z-statistic, we run the bootstrap on the same test 1000 times and receive a matrix with the strategies' z-statistic distributions. For each strategy, the mean z-stat is subtracted from the vector of observed strategy z-values, from where the resulting vector is maximized, and strategy selected.

The other suggested improvement is to decrease the test's sensibility to the impact of poor performing strategies. In our context, this refers to the risk of making strategies appear better than they are, due to the comparison to several bad performing strategies. The suggested solution is to exclude strategies with a test statistic z_k below the threshold value A, where $A = -\sqrt{2ln(ln(n)}$, and n refers to sample size. In our SPA test of US data for Sharpe ratios as performance measures, we reveal 8 out of 135 strategies with a test statistic below the threshold value:

$$A_{USA} = -\sqrt{2ln(ln(618)} = -1.93.$$

The strategies below A are then excluded before the SPA test is conducted a second time.

The maximized bootstrapped strategy is then compared with the best observed strategy, where the p-value is computed in the same way as demonstrated in the section of Sharpe ratio in WRC in equation 23.

Compared with WRC, the Sharpe ratio differences between the strategies and benchmark are now computed in a z-test for Sharpe ratio and t-test for FFC 4-factor alpha. In other words, as pointed out initially, the Sharpe ratio differences are first normalized to compare strategy portfolios at equal terms.

For the SPA test with 4-factor alpha as the performance measure, the procedure virtually follows the one for Sharpe ratio. The exception mainly lies in the computation of alphas, with the inclusion of multiple regressions, where the market, along with factors for firm characteristics, determines the perceived performance, rather than the risk-rewarding ratio. When strategy and the benchmark alphas are regressed, the outperformances against benchmark are expressed by a t-test. We get an observed maximized vector of strategy outperformances, which in the end is compared with the maximization of the vector with bootstrapped strategy alphas, subtracted for observed strategy alphas to comply with the null hypothesis, in the same way as for WRC.

5 Empirical Results

In this section, we will present the results of our empirical study. First, we show the Sharpe ratios (Table 3) and 4-factor alphas (Table 4) for all strategies along with their associated p-values. Next, we will examine the correlation matrix to reveal if there are high correlations that potentially can affect the z-tests used in the individual tests and the SPA test (Table 5). Third, we will present the results from the WRC for Sharpe ratios (Table 6) and 4-factor alphas (Table 7), followed by the results from the Sharpe ratios (Table 8) and 4-factor alphas (Table 10) at the SPA test. At last, we summarize all results as a basis for further discussion (Table 13).

In the WRC and SPA tests, our null hypothesis is that the best performing strategy will not outperform the 1/N benchmark, as presented in Section 4. In addition, we have for both test methodologies conducted an additional hypothesis test where our null hypothesis is that the second best, third best, and fourth-best strategy does not outperform the benchmark.

The motivation behind the additional hypotheses is twofold. First, we want to get an impression of if there are any differences in the number of significant strategies when using WRC versus the SPA test, and between the US- and Norwegian data. Second, we want to see if any strategies are repeatedly significant in the different tests and different sources of data. The reason for the exact number of 4, is a result of trial and error through testing. When choosing the 4 best strategies, we observe that this number of times covers the significant p-values and best illustrates our motivation.

5.1 Traditional Hypothesis Testing

In previous research, traditional methodology applied to reveal potential significant strategy performances is parametric approaches such as the pairwise z- or t-tests between the strategies and the benchmark. These tests are conducted by first comparing the individual strategies performances in terms of Sharpe ratio, alpha, or other performance measures applied. The strategy outperformances to benchmark are then normalized by dividing the outperformance on the strategy's and benchmark's standard deviations. The computed z-statistic will then be compared to a distribution with assumptions towards normality and independency. For illustration purposes, we first present the results of our significant test conducted the traditional way. Within each dataset, we employ the parametric z-test, similar to previous research such as DeMiguel et al. (2009), Kirby and Ostdiek (2012), and others.

Market	Dataset	Naive	RP	ERC	MDP	VolTim	RRTim	MinVar	MVP	MaxDec	RiskEff
USA	Size	0.41	0.42	0.42	0.40	0.43	0.43	0.41	0.43	0.39	0.41
			(0.108)	(0.123)	(0.666)	(0.151)	(0.354)	(0.484)	(0.399)	(0.806)	(0.658)
	B/M	0.48	**0.49**	0.49	0.44	**0.51**	0.52	0.49	0.46	0.43	0.46
			(0.038)	(0.089)	(0.956)	(0.048)	(0.231)	(0.466)	(0.643)	(0.985)	(0.860)
	OP	0.39	**0.40**	**0.40**	0.36	**0.43**	0.41	**0.50**	0.40	0.33	0.38
			(0.000)	(0.000)	(0.943)	(0.000)	(0.337)	(0.001)	(0.398)	(0.996)	(0.798)
	Inv	0.46	**0.47**	**0.47**	0.46	**0.50**	0.52	0.50	0.45	0.44	0.44
			(0.000)	(0.000)	(0.531)	(0.002)	(0.125)	(0.154)	(0.526)	(0.771)	(0.875)
	Mom	0.37	**0.39**	**0.39**	0.38	**0.43**	**0.52**	**0.44**	**0.55**	0.35	**0.44**
			(0.002)	(0.006)	(0.345)	(0.003)	(0.016)	(0.035)	(0.010)	(0.715)	(0.004)
	STR	0.40	0.41	0.40	0.33	0.42	**0.51**	0.41	0.48	0.32	0.35
			(0.088)	(0.230)	(0.999)	(0.097)	(0.030)	(0.450)	(0.120)	(0.999)	(0.994)
	LTR	0.48	**0.49**	**0.49**	0.45	**0.51**	0.53	0.53	0.49	0.44	0.47
			(0.031)	(0.029)	(0.895)	(0.045)	(0.177)	(0.089)	(0.433)	(0.932)	(0.682)
	ACC	0.41	**0.42**	**0.42**	0.42	**0.46**	0.47	**0.50**	0.45	0.40	**0.44**
			(0.000)	(0.000)	(0.229)	(0.001)	(0.111)	(0.022)	(0.241)	(0.680)	(0.033)
	Mkt B	0.41	**0.44**	**0.45**	0.46	**0.51**	0.41	0.54	0.33	0.39	**0.45**
			(0.001)	(0.001)	(0.109)	(0.005)	(0.504)	(0.058)	(0.858)	(0.736)	(0.042)
	NSI	0.35	**0.37**	**0.37**	**0.41**	**0.40**	**0.48**	0.40	**0.48**	**0.40**	0.37
			(0.001)	(0.001)	(0.010)	(0.002)	(0.009)	(0.173)	(0.021)	(0.017)	(0.225)
	ResVar	0.35	**0.40**	**0.39**	0.27	**0.48**	**0.45**	**0.51**	0.44	0.15	0.35
			(0.000)	(0.000)	(0.996)	(0.000)	(0.048)	(0.014)	(0.124)	(1.000)	(0.551)
	Var	0.36	**0.41**	**0.41**	0.30	**0.51**	0.40	**0.52**	0.31	0.16	0.37
			(0.000)	(0.000)	(0.936)	(0.001)	(0.261)	(0.035)	(0.746)	(1.000)	(0.343)
	E/P	0.49	0.50	0.49	0.48	0.50	0.54	0.46	0.48	0.46	0.47
			(0.237)	(0.543)	(0.739)	(0.363)	(0.175)	(0.798)	(0.575)	(0.853)	(0.866)
	CF/P	0.49	0.49	0.49	0.47	0.49	0.55	0.47	0.54	0.45	0.47
			(0.173)	(0.328)	(0.763)	(0.299)	(0.098)	(0.688)	(0.201)	(0.877)	(0.736)
	DivY	0.48	**0.49**	**0.49**	0.46	**0.53**	0.44	0.52	0.39	0.45	0.48
			(0.004)	(0.013)	(0.631)	(0.007)	(0.751)	(0.217)	(0.891)	(0.795)	(0.424)
Norway	B/M	0.77	0.78	0.78	0.76	0.79	0.67	0.77	0.63	0.76	0.79
			(0.256)	(0.205)	(0.568)	(0.301)	(0.915)	(0.517)	(0.951)	(0.619)	(0.362)
	Size	1.32	1.30	1.34	**1.44**	1.27	**1.47**	1.27	**1.50**	**1.47**	1.32
			(0.901)	(0.050)	(0.022)	(0.870)	(0.030)	(0.712)	(0.047)	(0.008)	(0.444)
	Mom	0.78	0.77	0.77	0.80	0.74	0.80	0.70	0.81	0.82	0.78
			(0.870)	(0.676)	(0.278)	(0.882)	(0.350)	(0.871)	(0.356)	(0.202)	(0.441)
	Spread	1.07	1.08	**1.11**	**1.24**	1.08	**1.24**	1.12	**1.29**	**1.22**	1.14
			(0.219)	(0.002)	(0.004)	(0.459)	(0.036)	(0.269)	(0.032)	(0.009)	(0.057)

Table 3: This table reports the annualized Sharpe ratio for all strategies listed in Table 2 for each dataset reported in Table 1. In brackets is the p-value of the difference between the naive strategy and the optimized strategies. Statistically significant values at 5% level is reported in bold text.

Out of the 135 p-values from the 9 strategies and 15 datasets in US data in Table 3, we observe that 48 p-values (35.6%) are significant at the 5% level. Further, all strategies are significant in some set of data, where at the same time, all strategies are insignificant in some other set of data. For Norwegian data, we have 10 out of 36 significant p-values (22.7%). Similar to the US data, we do not reveal any pattern in which strategies that are significant and not.

For the list of p-values and 4-factor alphas in Table 4, the tendencies of significant

Market	Dataset	Naive	RP	ERC	MDP	VolTim	RRTim	MinVar	MVP	MaxDec	RiskEff
USA	Size	-0.09	-0.03	**-0.02**	-0.22	0.14	0.56	0.23	0.98	-0.38	-0.07
			(0.059)	(0.042)	(0.615)	(0.060)	(0.251)	(0.224)	(0.188)	(0.725)	(0.471)
	B/M	0.11	0.20	0.19	-0.25	0.33	-0.26	0.02	-1.30	-0.51	-0.01
			(0.057)	(0.069)	(0.827)	(0.107)	(0.700)	(0.564)	(0.938)	(0.933)	(0.661)
	OP	-0.28	**-0.12**	**-0.13**	-0.51	**0.19**	-0.19	**1.04**	-0.40	-0.83	-0.49
			(0.000)	(0.001)	(0.774)	(0.001)	(0.453)	(0.002)	(0.550)	(0.948)	(0.796)
	Inv	0.49	**0.57**	0.56	0.19	0.67	0.44	0.46	-0.61	0.15	0.27
			(0.045)	(0.052)	(0.862)	(0.113)	(0.524)	(0.524)	(0.889)	(0.879)	(0.807)
	Mom	0.38	0.39	0.31	0.04	0.42	-0.08	0.40	0.44	-0.36	0.52
			(0.475)	(0.743)	(0.767)	(0.442)	(0.718)	(0.489)	(0.475)	(0.920)	(0.340)
	STR	0.09	0.08	0.04	-0.98	0.09	**1.70**	-0.20	1.48	-1.01	-0.81
			(0.576)	(0.780)	(0.998)	(0.512)	(0.036)	(0.718)	(0.088)	(0.997)	(0.998)
	LTR	0.43	0.49	0.49	0.10	0.57	0.70	0.76	0.10	-0.02	0.39
			(0.209)	(0.200)	(0.799)	(0.266)	(0.368)	(0.269)	(0.636)	(0.838)	(0.556)
	ACC	0.66	**0.78**	**0.78**	0.56	**1.03**	0.80	1.26	0.58	0.38	**1.22**
			(0.008)	(0.008)	(0.652)	(0.018)	(0.424)	(0.138)	(0.535)	(0.849)	(0.014)
	Mkt B	0.09	**0.30**	0.31	0.55	0.75	-0.55	1.20	-2.11	0.13	0.40
			(0.047)	(0.062)	(0.195)	(0.064)	(0.737)	(0.101)	(0.965)	(0.464)	(0.167)
	NSI	-0.51	**-0.41**	**-0.39**	0.00	-0.20	**1.33**	-0.20	**1.59**	-0.07	-0.41
			(0.044)	(0.047)	(0.058)	(0.071)	(0.012)	(0.308)	(0.014)	(0.075)	(0.352)
	ResVar	-0.39	**-0.08**	**-0.12**	-2.00	**0.58**	0.08	**0.97**	0.08	-3.44	-0.83
			(0.016)	(0.035)	(0.999)	(0.012)	(0.293)	(0.039)	(0.325)	(1.000)	(0.902)
	Var	-0.55	**-0.13**	**-0.17**	-1.81	**0.64**	-0.77	0.85	-2.25	-3.30	-0.57
			(0.010)	(0.027)	(0.978)	(0.018)	(0.594)	(0.071)	(0.942)	(1.000)	(0.516)
	E/P	0.67	0.64	0.64	0.78	0.57	0.59	0.10	-0.22	0.78	0.56
			(0.697)	(0.766)	(0.385)	(0.740)	(0.547)	(0.852)	(0.843)	(0.388)	(0.658)
	CF/P	0.49	0.49	0.47	0.26	0.44	0.54	0.25	0.28	0.27	0.32
			(0.555)	(0.689)	(0.721)	(0.641)	(0.473)	(0.675)	(0.592)	(0.703)	(0.742)
	DivY	0.35	**0.47**	0.48	0.27	**0.82**	-0.19	0.82	-0.83	0.17	0.45
			(0.039)	(0.084)	(0.565)	(0.037)	(0.757)	(0.263)	(0.891)	(0.648)	(0.374)
Norway	B/M	-5.51	**-5.15**	**-4.97**	-4.08	**-3.93**	-7.00	**-1.85**	-6.53	-4.47	**-2.94**
			(0.016)	(0.004)	(0.051)	(0.009)	(0.826)	(0.007)	(0.699)	(0.102)	(0.000)
	Size	0.05	-0.26	0.31	**3.00**	-0.53	**5.33**	0.82	**7.57**	**3.70**	0.96
			(0.957)	(0.110)	(0.001)	(0.788)	(0.000)	(0.290)	(0.000)	(0.000)	(0.080)
	Mom	-6.14	-6.17	-6.15	-5.67	-6.05	-5.39	-5.23	-5.51	-5.46	-5.40
			(0.565)	(0.508)	(0.310)	(0.441)	(0.301)	(0.254)	(0.364)	(0.244)	(0.170)
	Spread	-3.09	-3.12	**-2.75**	**-0.16**	-3.35	**0.70**	-1.93	**2.94**	**-0.42**	**1.36**
			(0.579)	(0.044)	(0.001)	(0.688)	(0.005)	(0.174)	(0.000)	(0.003)	(0.006)

Table 4: This table reports the annualized Fama/French Carhart 4-factor alpha values for all strategies listed in Table 2 for each dataset reported in Table 1. In brackets is the p-value of the difference between the naive strategy and the optimized strategies. Statistically significant values at 5% level is reported in bold text.

p-values follow the table for Sharpe ratios. For US data, we have significant p-values for 25 out of 135 strategies (18.5%), while for Norwegian data, we have significant values for 15 out of 36 strategies (41.7%). Similar to the table of Sharpe ratios, the pattern of significant p-values is characterized by randomness.

The results illustrate several issues with regard to this testing methodology. First, there are obvious challenges in the interpretation of significant p-values for each of the strategies when they only appear in some of the datasets, while not in others, and where the significant values follow no pattern in the direction of some particular strategy being significant more often than others. In addition, we see that, for instance, the risk parity portfolio (RP), which is significant in 10 out of 15 datasets in US data, is not significant in none of the datasets for Norwegian data.

Another issue with regard to inference is the data-mining bias. When testing 9 strategies in 15 datasets as for US data, we have in total 135 p-values. We know that, when testing one strategy against the benchmark, the chance of falsely rejecting the null hypothesis at a 5% significance level is $1 - (1 - 0.05)^1 = 5\%$. When we are testing 135 strategies, the probability of rejecting the null due to type 1 error in at least one of the strategies is $1 - (1 - 0.05)^{135} = 99.9\%$.

Hence, in our testing methodology, we consider this phenomenon by adjusting the performances to comply with the null hypothesis, by subtracting the observed outperformances from the bootstrapped outperformances, in the application of the SPA test and WRC.

Another potential explanation of the many significant results when conducting the z-test is the use of the correlation in the z-test's denominator.

If the correlations between the strategies' and benchmark's excess returns are high, we can see that the denominator will be small, resulting in a high z-statistic, which again provides small p-values. In addition to the p-values, we can in Table 3, and Table 4 observe that many of the strategies' Sharpe ratios and 4-factor alphas are just marginally higher than for benchmark. Despite the marginal differences in performances, the tables reveal several significant results. Economically, the inference of a Sharpe ratio difference of $0.40 - 0.39 = 0.01$ as for our third dataset, Operating Profitability, is that their performances are approximately equal. Further, in the list of 4-factor alphas in our first dataset Size, the difference between naive and ERC is -0.09 versus -0.03. With a difference of only 0.06, the p-value is still highly significant. On the contrary, in the same dataset, the MVP portfolio has the highest alpha of 0.98. The economic distance to benchmark is higher, but now, this strategy's outperformance is not statistically significant. To ensure that this phenomenon is not the case behind our significant results, we need to consider the correlation matrix between the strategies' and benchmark's excess returns within the different datasets. This is crucial with respect to the inferences based on the SPA test, due to the normalization of outperformances in this methodology, which is conducted by the parametric z-test.

Market	Dataset	RP	ERC	MDP	VolTim	RRTim	MinVar	MVP	MaxDec	RiskEff
USA	Size	0.9994	0.9993	0.9745	0.9901	0.9106	0.9048	0.8583	0.9788	0.9946
	B/M	0.9996	0.9996	0.9825	0.9962	0.9395	0.9621	0.9001	0.9810	0.9913
	OP	0.9997	0.9997	0.9896	0.9970	0.9350	0.9718	0.9035	0.9865	0.9935
	Inv	0.9995	0.9995	0.9920	0.9952	0.9260	0.9626	0.9002	0.9912	0.9932
	Mom	0.9981	0.9982	0.9769	0.9853	0.8670	0.9548	0.8356	0.9725	0.9817
	STR	0.9993	0.9994	0.9866	0.9937	0.9157	0.9673	0.8947	0.9852	0.9904
	LTR	0.9993	0.9994	0.9845	0.9935	0.9213	0.9617	0.8919	0.9788	0.9907
	ACC	0.9997	0.9997	0.9925	0.9950	0.9366	0.9527	0.9065	0.9924	0.9938
	Mkt B	0.9974	0.9966	0.9642	0.9629	0.8952	0.8207	0.8517	0.9788	0.9868
	NSI	0.9996	0.9994	0.9870	0.9940	0.9222	0.9470	0.8956	0.9891	0.9927
	ResVar	0.9969	0.9969	0.9739	0.9647	0.9106	0.8657	0.8645	0.9683	0.9875
	Var	0.9954	0.9948	0.9557	0.9416	0.9019	0.8012	0.8544	0.9575	0.9861
	E/P	0.9998	0.9998	0.9834	0.9973	0.9406	0.9619	0.9025	0.9797	0.9922
	CF/P	0.9998	0.9998	0.9830	0.9972	0.9296	0.9642	0.8960	0.9806	0.9916
	DivY	0.9993	0.9985	0.9579	0.9877	0.9138	0.9022	0.8673	0.9682	0.9885
Norway	B/M	0.9989	0.9982	0.9686	0.9812	0.9179	0.9039	0.8776	0.9739	0.9741
	Size	0.9983	0.9977	0.9424	0.9715	0.9001	0.8866	0.8352	0.9413	0.9786
	Mom	0.9988	0.9987	0.9686	0.9842	0.9274	0.9280	0.8912	0.9676	0.9778
	Spread	0.9988	0.9977	0.9416	0.9842	0.8708	0.9111	0.7984	0.9429	0.9753

Table 5: This table reports the correlation for all optimized strategies listed in Table 2 for each dataset reported in Table 1 to the naive strategy.

Table 5 reports the correlation matrix. The general impression is that an overwhelming majority of strategy-benchmark correlations are high and, in most cases, close to 1. This implies that the results from any z-test used, as for the SPA test, the inferences must be cautious with regard to the correlation of the best performing strategies to benchmark.

5.2 White's Reality Check

Table 6 presents the results from White's reality check for Sharpe ratios in US- and Norwegian data. As we can observe, none of the four best strategy outperformances are significant at a 5% level in either the US- or Norwegian data. The best performing

#	Dataset - Strategy USA	pvalue	Dataset - Strategy Norway	pvalue
1	Momentum - MVP	0.086	Spread - MVP	0.061
2	Variance - MinVar	0.173	Size - MVP	0.135
3	Residual Variance - MinVar	0.180	Spread - MDP	0.187
4	Momentum - RRTim	0.207	Spread - RRTim	0.192

Table 6: This table reports the 4 best strategies in respectively US and Norwegian market in terms of Sharpe ratio from White's reality check.

strategy in US-data is the MVP portfolio in the dataset sorted by Momentum with a p-value of 8.6%. In Norwegian data, the best performing strategy is the MVP portfolio in the dataset sorted by Spread, with a p-value of 6.1%

#	Dataset - Strategy USA	pvalue	Dataset - Strategy Norway	pvalue
1	Net Share Issues - MVP	0.279	Size - MVP	0.004
2	Net Share Issues - RRTim	0.410	Spread - MVP	0.016
3	Short-Term Reversal - RRTim	0.557	Size - RRTim	0.044
4	Variance - MinVar	0.684	Spread - RRTim	0.177

Table 7: This table reports the 4 best strategies in respectively US and Norwegian market in terms of Fama/French Carhart 4 Factor alpha values from White's reality check.

Table 7 shows the four highest outperformances against the benchmark for the US- and Norwegian data with respect to 4-factor alpha. For US data, we observe that none of the four best strategies are significant. However, for Norwegian data, we have three strategies from three datasets, which are significant at the 5% level. The best performing significant strategy is the MVP portfolio in the dataset sorted by size, with a p-value of 0.4%.

5.3 SPA test

In the SPA test, when using Sharpe ratios as performance measures, Table 8 reports significant p-values for the four best performances in US data, while none for the Norwegian data. Due to the cautiousness with regard to our earlier demonstration of the z-test applied in the SPA test, we control for the correlations of the significant strategies. In the correlation matrix, we can observe the correlations between the strategies and

#	Dataset - Strategy USA	pvalue	Dataset - Strategy Norway	pvalue
1	Operating Profitability - RP	0.005	Spread - ERC	0.053
2	Operating Profitability - ERC	0.007	Spread - MDP	0.079
3	Operating Profitability - VolTim	0.009	Size - MaxDec	0.155
4	Variance - RP	0.013	Spread - MaxDec	0.167

Table 8: This table reports the 4 best strategies in respectively US and Norwegian market in terms of Sharpe ratio from superior predictive ability test.

the benchmark. In addition, we revisit the table of strategies Sharpe ratios to get an impression of the economic differences between the strategies and benchmark.

#	Dataset - Strategy	Correlation	Sharpe ratio difference
1	Operating Profitability - RP	0.9997	0.01
2	Operating Profitability - ERC	0.9997	0.01
3	Operating Profitability - VolTim	0.9970	0.04
4	Variance - RP	0.9954	0.05

Table 9: This table reports the correlation ratio between the naive strategy and 4 best active strategies in terms of Sharpe ratio, and their difference in Sharpe ratio to the naive strategy.

As Table 9 reveals, despite that the strategies p-values are highly significant, we can observe that their economic performances are approximately equal. Further, we can conclude those correlations are all extremely high and close to 1, making the z-values high and p-values low.

#	Dataset - Strategy USA	pvalue	Dataset - Strategy Norway	pvalue
1	Operating Profitability - RP	0.005	Size - MVP	0.001
2	Operating Profitability - ERC	0.007	Size - MaxDec	0.013
3	Operating Profitability - VolTim	0.009	Size - RRTim	0.015
4	Variance - RP	0.012	Spread - MVP	0.045

Table 10: This table reports the 4 best strategies in respectively US and Norwegian market in terms of Fama/French Carhart 4 Factor alpha values from superior predictive ability test.

For the 4-factor alphas in the SPA test, Table 10 presents significant p-values for all of the four best strategies in both US- and Norwegian data. For the same reason as for Sharpe ratios, we consider the correlation matrix and the strategies 4-factor alpha differences to benchmark.

#	Dataset - Strategy	Correlation	Alpha difference
1	Operating Profitability - RP	0.9997	0.21
2	Operating Profitability - ERC	0.9997	0.21
3	Operating Profitability - VolTim	0.9970	0.63
4	Residual Variance - RP	0.9969	0.62

Table 11: This table reports the correlation ratio between the naive strategy and 4 best active strategies in terms of 4-factor alpha values, and their difference in alpha values to the naive strategy.

Table 11 present the statistics from US data. As for the Sharpe ratios, again, we observe extreme correlations and marginal economic differences.

To avoid deceptions in inferences, we will, therefore, be emphasizing the results from White's reality check, which as derived earlier, on the contrary to the SPA test, only consider Sharpe ratio differences in the comparison of strategies versus the benchmark.

5.4 Summary of Results

For the null hypothesis that the best performing strategy does not outperform the benchmark, Table 12 summarize the following results:

	USA		Norway	
	Sharpe	Alpha	Sharpe	Alpha
WRC	0.086	0.279	0.061	0.004
SPA	0.005	0.005	0.053	0.001

Table 12: This table reports a summary of the different tests, and lists the p-value from the best strategy in each test

In total, the best performing strategy has significant outperformance to benchmark in 3 out of 4 times when conducting the SPA test in both performance measures in US data and 1 out of 2 times in Norwegian data. For the WRC, we have 1 out of 4 significant results; this is from the 4-factor alpha in Norwegian data.

For the hypotheses that the second, third, and fourth-best performing strategies outperforms benchmark, Table 13 reports the following results.

Out of the 32 p-values, 15 are significant at the 5% level. 3 out of 16 for the WRC,

		USA		Norway	
		Sharpe	Alpha	Sharpe	Alpha
WRC	1	0.086	0.279	0.061	0.004
	2	0.173	0.410	0.135	0.016
	3	0.180	0.556	0.187	0.044
	4	0.207	0.684	0.192	0.177
SPA	1	0.005	0.005	0.053	0.001
	2	0.007	0.007	0.079	0.013
	3	0.009	0.009	0.155	0.015
	4	0.013	0.012	0.167	0.045

Table 13: This table summarizes the p-values of the 4 best strategies from each test

and 12 out of 16 for the SPA test.

However, we have shown how high correlations between strategy- and benchmark's excess returns affect the p-values when conducting the SPA test. A lower economic difference with higher correlation was statistically significant, where a higher economic difference with a lower correlation was not. In addition, we have presented the correlation matrix between all strategies and benchmark for every data sets, revealing very high correlations in general. Summarized, these drawbacks urge caution in inferences based on the SPA test in this context. We also have significant alphas for the WRC in Norwegian data when using alpha as a performance measure, while none are significant when using Sharpe ratio, and none are significant in US data for both performance measures. These results will be further discussed in Section 6.

6 Discussion

Initially, in Section 5, we presented the p-values from the strategies' individual tests against the benchmark in all datasets. Table 3 reports the p-values for strategy Sharpe ratios, while Table 4 presents the FFC 4-factor alphas. The tables illustrate which results we achieve by conducting the traditional individual outperformance tests. As we can see from Table 3 for US data, the p-values report significant results for approximately 36% of the strategy outperformances in terms of Sharpe ratio. When conducting the hypothesis test by WRC for the same performance measure, none of the four best performing strategies have significant results. These contrary results illustrate the purpose of WRC and SPA-test with regard to data-mining bias as derived in section 4. In addition, we see that the significant p-values in the one set of data are not significant in the next. In other words, the significant results reveal no pattern in the strategy performances, and that the strategy performances are depending on the characteristics of each set of data, rather than the strategies themselves.

From the SPA test, we can observe that the test produces several significant p-values in both US- and Norwegian data for both performance measures used. However, most of these strategies have an approximately equal economic performance to benchmark, with respect to Sharpe ratios and FFC 4-factor alphas. In Table 5, we revealed highly correlated excess returns between strategies and benchmark. Further, in Section 5, we demonstrated how these correlations affect the z-values, which again produce low p-values. The conclusion based on these findings is that the SPA test in our thesis produces several statistically significant performances without economic significance. As a consequence, our inferences will mainly be drawn based on WRC, which in the computation of strategy outperformances considers only the Sharpe ratio difference without the correlation factor affecting the results.

Our thesis follows the numerous attempts to explore strategy outperformances against some benchmark. In Section 2, we learned how the performances of actively managed portfolios against the naive benchmark are highly debated. DeMiguel et al. (2009) find that none of the strategies based on Markowitz's mean-variance framework outperforms benchmark consistently in terms of Sharpe ratio and two other performance measures. In the wake of these findings, several of the following studies claim the opposite (Kritzman et al. (2010), (Tu and Zhou (2011) and (Kirby and Ostdiek (2012)). In defense of the findings in DeMiguel et al. (2009), Zakamulin (2017) points out several commonalities between the most relevant previous studies, which claim the existence of outperforming strategies. First, the studies simulate strategy outperformances in the same source of data, using Kenneth French's online data library of US data. Second, the strategy performances are often evaluated in the light of their Sharpe ratios, which implies not controlling for the anomaly characteristics. When the Sharpe ratio is used in the evaluation of performance, we know that the ratio is computed as the relation between excess returns and their standard deviation. Further, in Section 2, we learn how low-volatility anomalies facilitated tendencies for low-volatility firms to yield higher returns, in opposite to the intuitivism of rewarded risk. This results in portfolios of low-volatility assets, which provides higher Sharpe ratios due to both increased returns and less risks. However, the portfolio performances in such cases, are now explained by the anomaly affects and not by the strategies themselves. In addition to low-volatility anomaly, we know from Section 2, the existence of several anomaly effects, which are not captured when Sharpe ratio is the performance measure of choice. To deal with these effects, we employed the FFC 4-factor model as an additional performance measure in our thesis. Another essential commonality in previous literature is the use of the individual testing methodology. In section 5, we observe the results when the individual z-test is applied as the basis for the hypothesis tests and concludes that many significant p-values arise due to data-mining bias. To control for such bias, we conduct White's reality check and its extended SPA test.

From our 15 US datasets and 4 Norwegian datasets, we have no consistency in any of the strategies' outperformances against the 1/N portfolio in terms of traditional testing. When WRC is applied as methodology, the test reports non-significant p-values for the four best strategy performances for US data. For Norwegian data, WRC reports that the 3 best performing strategies have significant p-values against the benchmark when the FFC 4-factor alpha is used as the measure of performance. These significant results are not supported by the test using Sharpe ratio. A possible explanation could be that the Norwegian market is not as efficient as the US market, in the sense that all available information is priced into the assets. If this is the case, the theory of market efficiency enables the possibility to beat the market. Hence, by FFC 4-factor alpha as the excess return to the market, this inefficiency is accounted for. For the Sharpe ratio, on the other hand, the returns are based on the assumptions of CAPM, which assumes a highly efficient market.

These findings open for further questions and future efforts within research. How do the optimizing strategies perform in different markets, employing combined tests and factor models? In addition, it could be interesting to reveal if the different markets are characterized by different anomalies. For the Norwegian market, Næs, Skjeltorp, and Ødegaard (2009) claim that the returns are mostly affected by size and liquidity. An interesting study could be to test the strategies in the Norwegian market by adding a factor for liquidity. However, despite the significant results for 3 strategies in Norwegian data, the same strategies are not significantly superior to benchmark in the data from the US. Even if the strategies perform better in Norwegian data, for a certain period, we cannot generalize these results to a conclusion of strategy superiority in general. Hence, in the light of superior performance dedicated to the strategies themselves, our findings still support the conclusion of DeMiguel et al. (2009), Zakamulin (2017), Hsu et al. (2018) and (Yang et al. (2019).

7 Conclusion

In the wake of the mean-variance framework of Markowitz (1952), a considerable effort has been invested in the search for solutions to reduce the gap between estimated and out-of-sample parameters. Despite the remarkable development, the debate on whether optimal portfolios are superior to the 1/N-benchmark is still ongoing. However, due to the steady progress in testing methodology, we are now able to get more reliable results in the testing of hypotheses about portfolio performances. Motivated by the great work from previous studies and these new testing opportunities, we want to replicate previous research, using the extended testing methodology. Our goal is to test whether optimized portfolio strategies outperform the 1/N portfolio. To deal with the critiques regarding the results of earlier findings, we account for potential data-mining bias in individual testing using WRC and SPA test. Further, as an extension of the Sharpe ratio as a performance measure, we are adopting the FFC 4-factor alpha, to meet with the challenges of data anomaly. Last, we have extended previous studies by the inclusion of Norwegian data, which provides a broader basis for inferences than just the simulation of strategies in US data alone.

When using the SPA test, our results report that the four best strategies are statistically significant in terms of both Sharpe ratio and FFC 4-factor alpha in US data. For Norwegian data, the 4 best performing strategies are significant in terms of FFC 4-factor alpha, but not for Sharpe ratio. Still, despite the statistical significance, the performances are not economically significant. Based on the SPA test, our conclusion is in accordance with the null hypothesis. When using WRC in our hypothesis tests, we find that the four best strategies do not outperform the benchmark in US data in either of the performance measures. For Norwegian data, we find that 3 of the 4 best strategies outperform benchmark significantly when using the FFC 4-factor alpha, but not in terms of Sharpe ratio. Our conclusion is that the 3 best strategies can outperform the benchmark in Norwegian data, where one possible explanation can be due to market efficiency. Despite the significant results for Norwegian data, we conclude that optimizing strategies' superiority is still yet to be proved in general.

References

Amenc, N., Goltz, F., Martellini, L., and Retkowsky, P. (2011). "Efficient indexation: an alternative to cap-weighted indices", *Journal of Investment Management*, *9*(4), 1–23.

Barry, C. B. (1974). "Portfolio analysis under uncertain means, variances, and covariances", *Journal of Finance*, *29*(2), 515–522.

Black, F., Jensen, M. C., and Scholes, M. (1972). "The capital asset pricing model: some empirical tests", *Studies in the theory of capital markets*, *81*(3), 79–121.

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

Chaves, D., Hsu, J., Li, F., and Shakernia, O. (2012). "Efficient algorithms for computing riskparity portfolio weights", *Journal of Investing*, *21*(3), 150–163.

Choueifaty, Y. and Coignard, Y. (2008). "Toward maximum diversification", *Journal of Portfolio Management*, *35*(1), 40–51.

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

Clarke, R., De Silva, H., and Thorley, S. (2011). "Minimum-variance portfolio composition", *Journal of Portfolio Management*, *37*(2), 31–45.

Clarke, R. G., de Silva, H., and Thorley, S. (2006). "Minimum-variance portfolios in the U.S. equity market", *Journal of Portfolio Management*, *33*(1), 10–24.

DeMiguel, V., Garlappi, L., and 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. F. and 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., and Ostdiek, B. (2001). “The economic value of volatility timing”, *Journal of Finance*, *56*(1), 329–352.

Fleming, J., Kirby, C., and Ostdiek, B. (2003). “The economic value of volatility timing using ”realized” volatility”, *Journal of Financial Economics*, *67*(3), 473–509.

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

Haugen, R. A. and Heins, A. J. (1972). “On the evidence supporting the existence of risk premiums in the capital market”, Working Paper.

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

Jegadeesh, N. and Titman, S. (1993). “Returns to buying winners and selling losers: implications for stock market efficiency”, *Journal of Finance*, *48*(1), 65–91.

Kirby, C. and Ostdiek, B. (2012). “It’s all in the timing: simple active portfolio strategies that outperform naive diversification”, *Journal of Financial and Quantitative Analysis*, *47*, 437–467.

Kritzman, M., Page, S., and Turkington, D. (2010). “In defense of optimization: the fallacy of 1/N”, *Financial Analysts Journal*, *66*(2), 31–39.

Ledoit, O. and Wolf, M. (2004). “Honey, I shrunk the sample covariance matrix”, *Journal of Portfolio Management*, *30*(4), 110–119.

Lintner, J. (1965). “Security prices, risk, and maximal gains from diversification”, *Journal of Finance*, *20*(4), 587–615.

Maillard, S., Roncalli, T., and Teiletche, J. (2010). "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.

Masters, T. (2018). *Assessing and improving prediction and classification: theory and algorithms in C.* Apress.

Merton, R. C. (1980). "On estimating the expected return on the market: an exploratory investigation", *Journal of Financial Economics*, *8*, 323–361.

Mossin, J. (1966). "Equilibrium in a capital asset market", *Econometrica: Journal of the econometric society*, *34*(4), 768–783.

Næs, R., Skjeltorp, J. A., and Ødegaard, B. A. (2009). "What factors affect the Oslo Stock Exchange?", Working paper 24, Norges Bank.

Plyakha, Y., Uppal, R., and Vilkov, G. (2014). "Equal or value weighting? Implications for asset-pricing tests", Working paper, SSRN.

Politis, D. N. and Romano, J. P. (1994). "The stationary bootstrap", *Journal of the American Statistical association*, *89*(428), 1303–1313.

Politis, D. N. and White, H. (2004). "Automatic block-length selection for the dependent bootstrap", *Econometric Reviews*, *23*(1), 53–70.

Qian, E. (2011). "Risk parity and diversification", *Journal of Investing*, *20*(1), 119–127.

Sharpe, W. F. (1964). "Capital asset prices: a theory of market equilibrium under conditions of risk", *The Hournal of Finance*, *19*(3), 425–442.

Treynor, J. L. (1962). "Jack Treynor's' toward a theory of market value of risky assets"', Manuscript.

Tu, J. and 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., and Wang, Q. (2019). “Tactical asset allocation on technical trading rules and data snooping”, *Pacific-Basin Finance Journal*, *57*, 101049.

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

8 Appendix

8.1 Reflection Note 1

Our master thesis is an replication and extension of previous research, where the purpose has been and to evaluate whether portfolio strategies based on the mean-variance framework of Markowitz (1952) have superior performance to the simple naive portfolio. In the previous literature this subject has been debated, where the different studies have claimed findings in opposite and spread directions. Our thesis is an extension due to the methodology adopted in our study. In the majority of previous research, the hypothesis test of the portfolio performances has been conducted by parametric testing methods, where each portfolio strategy is tested against some benchmark individually. In addition, the performances measures used has later been found to ignore different anomalies represented in the data used for simulation of strategies. Anomalies in this context refers to asset performances which are due to some other asset characteristics which are not captured by the applied performance measure, like for instance risk/volatility when using Sharpe ratio, or market returns using the CAPM alpha. In our thesis, we replicate previous research, by the adoption of methods which considers these earlier shortcomings.

Our thesis is conducted in a highly international context. Previous contextual studies has been provided from the academic environment from all over the world, with the American Harry Markowitz (Markowitz, 1952) as a sort of a founder of modern portfolio theory. Further through Victor Demiguel ((DeMiguel et al. (2009)) who later criticized Markowitz‘z mean-variance framework for lack of practical implementation, to Chinese Academics (Yang et al. 2019) which has provided empirical studies latest in the previous years. Portfolio theory is an international language, meaning, the formulas and theories is universal in the way they work, being relevant to the stock indexes of USA as the Oslo stock exchange. Meaning, the strategies do not consider the different indexes‘ nationality in their computation of portfolio weights. Another international aspect of portfolio theory as a technical branch of wealth management, is the investments on cross of borders. If we take the Norwegian pension fund for instance, the surplus from the oil industry is invested in 74 countries, taking part in global growth and wealth creation all over the world, while at the same time diversifying risk on behalf of its own economy (Norges Bank, 2020). I the wake of the increased globalization, portfolio management as an international field of research will follow as a natural cause, as the sharing of knowledge and flow of information is available as never before.

Another aspect of portfolio theory is innovation. In the wake of the critique against Markowitz‘s framework (1952), several attempts has been made to improve the frameworks issues with regard to estimation errors when predicting future portfolio performances. One early attempt was to estimate future performance by excluding asset returns in the computation of portfolio weights, and instead only consider their volatilities, as a more reliable source for estimation. Another attempt has been the implementations of methods to reduce the impact of potential estimation errors, as the Bayesian approach (Barry, 1974) and shrinkage methodology (Ledoit and Wolf (2004)). In addition, we have the inventions of all the different strategies either they are using just the variances or the hole mean-variance framework as base. They all have their own invented ways of allocating asset weights, based on detailed rules, aiming to achieve the same primary goals. Besides the portfolio strategies, we have seen how the hypothesis- and testing methodology has been subject to several pointed out issues and critiques towards lack of robustness as reliance on testing results. As a consequent, tedious attempts has been made to invent methodology that deals with the different problems in previous works. One example is the White's reality check (White (2000)), which arose as a critique against parametric testing with regard to data mining bias. The invention is now adopted by several new studies, including our own thesis. However, even though the methodology is improving, the strategies is increasing in sophistication, the methodology still have its drawbacks, and strategies are still yet to proven as superior alternatives to the more simplified diversifications. This challenge the next generations of researchers to still keep inventing and improving, in the spirit of the work and results which are already provided.

Portfolio theory is about techniques and methodology in how to preserve and create wealth to the benefit for companies, their employees, their families and their societies. In addition, we have seen how wealth in one country is diversified on cross of several countries (Norges Bank, 2020) which helps less wealthy societies in building up businesses, health care, infrastructure and a potentially safer and better quality of life. But we have also seen the other side of the scale. How people how lost their jobs, their homes, their health and safety, when those in charge of preserving wealth, preserved it on the cost of ethical behavior, where the personal enrichment was created by the exploitation of deregulations, natural- and human resources. Wealth creation can both ruin and built. The difference between the two opposites are depending on an increasingly and steady implementation of ethical standards, in the education of those who will be and in the regulation of those who are, given the responsibility to preserve wealth on behalf of others. These others can be the investors, but also those who are influenced by the different investments. With the pension fund as the world biggest fund, we have seen a development, where these considerations has evolved over time, as ethical awareness has been increasingly more emphasized. An awareness of how investments, not only should be considered by its short time growth, but where sustainability and ethical thinking has shown to create benefits for all in the long run.

The insights and knowledge of economics is the knowledge of society and peoples and the interactions between them. From the very early days of trades, where value was traded as fish for bread, we are now valuing production of products and services, often without the products and services being seen, and where value can be based on expectations of the future, without having a clue about what is around the next corner. The study of economics has provided a better overview of how world spins around, from the macro to micro perspective, with the insight that economy is the basis for all human life, either as the fish, or as the portfolio of assets which provides resources to the creation of products and services, in a constant development towards future.

8.2 Reflection Note 2

The main topic in this thesis is asset allocation and whether the straightforward, equally weighted portfolio or naive diversification is a good alternative to optimized portfolios. The naive diversification has been used as the benchmark for the optimized strategy. An empirical study has been conducted to determine whether any optimized strategy can outperform the naive diversification in the US and Norwegian markets. Our first performance measure was the Sharpe ratio, in which we in the SPA test found evidence for optimized strategies that are better than the naive diversification. However, that model has weaknesses that lead to significant results with very little difference in Sharpe ratios when the correlation between the returns is very high. That led to the use of the WRC test, which did not find any optimized strategy that outperforms the naive strategy at a 5% significance level.

The same procedure was done in the 4-factor alpha model, where we did not find evidence for strategies that outperformed the naive in the US market. For the Norwegian market, however, the story was different. In the end, we have signs of strategies that can beat the naive, but our main conclusion is that no strategy consistently outperforms the naive diversification in either of the markets.

This research field is very international as the theories and methodologies used in our thesis are based on international data. They are, however, very related to the US market, which is just one domestic market, but as that market attracts investors from the entire world, it is a very international market. To extend the research to a second market, we have added the Norwegian market. Both markets are tested over a specific limited period. But in terms of internationalism, it is essential to note that it is two different markets, and different factors may influence the returns. A model from the US market does not necessarily fit that well to the Norwegian market, which may explain the significant results. Still, to be consistent, we tested the same models for both markets.

The optimization of strategies evolves. We have not included any new strategy, but our combination of strategies and tests is relatively uncommon. There is still some innovation in the question if the optimized strategies can beat the naive diversification model. Since then, this discussion only started after DeMiguel et al. (2009) and has been a hot topic in the academic debate. Some of the strategies only evolved after the financial crisis. For example, the equal risk contribution model was formulated a few years after. Also, the continuous development in computer technology makes it possible for us to run a program that gives returns from these strategies. But, the only real innovation in our thesis is that we have included the USA and Norway in two different markets, and look at how the results correspond to each other.

Countless active fund management strategies are available to customers. But with no active strategy that is proved to beat the naive strategy, we can ask if it is necessary to pay for active portfolio management when the passive management yields the same results. There are, however, likely no funds that go strictly to one of the optimized strategies that we have studied, but our results still show that the naive strategy is better than it looks on the paper. And given how simple it is, it should get more attention in portfolio management. Still, there are likely few managers that want to give this advice as the portfolio fee would be minimal with this strategy versus an active one. And most of their customers are unknown of results like this, which creates an ethical discussion whether the portfolio management should go for what creates the highest profit for their fund or highest profit for their customers.