A Performance Evaluation of Norwegian Mutual Funds: Luck versus Skill

ANDRÉ JØRGENSEN
JONAS DRIVEKLEPP

SUPERVISOR
Valeriy Ivanovich Zakamulin

University of Agder, 2021
School of Business and Law
Department of Economics and Finance

A Performance Evaluation of Norwegian Mutual Funds: Luck versus Skill

André Jørgensen and Jonas Driveklepp

University of Agder

May 31, 2021

Abstract

In this master's thesis, we examine the performance of Norwegian mutual fund managers. Through a dataset of 107 Norwegian mutual funds' monthly returns from 1987-2019, we estimate fund managers' abnormal performance using Carhart (1997) four-factor model. First, we find that the managers cannot generate significant abnormal returns on an aggregate level. Further, we test the null hypothesis of zero performance in bootstrap approaches similar to Kosowski, Timmermann, Wermers and White (2006) and Fama & French (2010) to test whether the performance is is a result of luck or skills. We find no evidence of skill among the outperforming funds. However, we find evidence towards a lack of skill among the underperforming funds. Ultimately, we implement a new approach by Harvey & Liu (2020) for statistical testing under a new null hypothesis: A specific fraction of managers outperform or underperform the benchmark. The approach allows us to estimate the Type II error rate and the Test power. We find that the Test power is well below the recommended level. Hence, there might be some skillful managers in the Norwegian mutual fund market, though our test lacks the power to detect these. However, we find clear evidence towards the absence of skill under the assumption that a fraction of funds are underperforming.

Contents

List of Figures

List of Tables

1 Introduction

In this thesis, we evaluate the performance of actively managed Norwegian mutual funds. Investors may choose between actively managed funds that seek to generate a return in excess of a market benchmark or low-cost passively managed funds that replicate the benchmark index to match its result. If actively managed funds aspire to outperform their benchmark, the manager must display stock-picking skills. Such skills come at a price, and thus mutual fund managers require a premium for their services. These fees are deducted before evaluating the net asset value (NAV) of the fund, and as a consequence, the cost of active management directly affects the fund's net asset value. Furthermore, mutual fund managers must evaluate the market to identify and exploit incorrectly priced securities to generate abnormal returns. Hence, stock-picking skills are needed to create a portfolio of securities capable of beating the market.

A manager's performance is increasingly affected by external factors in the financial world due to increasing financial investment opportunities. These factors may affect the fund through market development and market behavior. These factors are further implemented through benchmark factor models to assess their impact on a fund's return. The idea is that, while the manager chooses the stocks, some companies display a return affected by market movement and are thus pre-disposed for growth. These attributes are considered excess of the manager's fundamental stock-picking skills and will serve as coefficients in the regression analysis of excess return. Existing literature aims to isolate the manager's performance through these factors and retain the performance in a single parameter, the alpha (α).

There are several approaches to evaluate the performance of mutual funds in previous literature. First, Jensen (1968) introduced Jensen‘s alpha, α, to measure fund managers' abnormal performance by excluding market factors. Further, Fama & French (1993) and Carhart (1997) added additional factors to increase the power in estimating the actual performance of funds. Next, Kosowski et al. (2006) introduced a bootstrap approach to separate skill from luck in the generated alphas. Their method tests the performance of managers individually under the null hypothesis of zero performance for all managers. Their results on the US fund market found that a minority of managers exhibit skill and generate returns to more than cover their costs.

Fama & French (2010) contributed to the test by Kosowski et al. (2006) to account for the potential correlation between fund returns and the possibility of survivorship bias in the sample. They use a similar dataset as Kosowski et al. (2006). However, their results suggest no skill among US mutual fund managers. Harvey and Liu (2020) further suggest that the absence of the rejection of the null hypothesis of zero performance might result from low Test power rather than a lack of skill. They introduce and test for the null hypothesis; a fraction of managers outperform or underperform. Which further allows for the estimation of the Type II error and the Test power.

Harvey & Liu (2020) identify modest evidence of skill among the US mutual fund managers with the same dataset.

Performance evaluation of the Norwegian mutual fund market frequently implements the approaches designed by Kosowski et al. (2006) and Fama & French (2010). For example, Sørensen (2011) implemented the approach of Kosowski et al. (2006) with modifications from the method of Fama & French (2010) onto a sample free of survivorship bias. He identified weak tendencies towards skill among the best performers. In contrast, he found clear signs of a lack of skill among managers. Further, the master's thesis of Braathe & Bjerke (2019) implemented Kosowski et al. (2006) onto a more recent dataset of funds' historical return. They identified no evidence of skill among the best-performing funds and found a clear indication of a lack of skill among the worst-performing funds. Previous studies have, to our knowledge, not yet implemented the recently developed approach of Harvey & Liu (2020) onto the Norwegian market.

Our research question is two-folded. **i)** Are Norwegian mutual fund managers able to beat their benchmark? Furthermore, **ii)** is their performance a result of luck or skill? These research questions are previously well-conducted in the performance evaluation of the Norwegian mutual fund market. However, there seems to be a lack of estimation of Type II error rate and Test power in the existing literature. Consequently, this master's thesis aims to supplement said literature by implementing the methods of Kosowski et al. (2006), Fama & French (2010) and Harvey & Liu (2020) onto the Norwegian mutual fund market.

We evaluate 107 Norwegian mutual funds' monthly returns in the time period 1987 to 2019. We require a minimum of 12 months of observations to consider possible survivorship bias in the data sample. First, we test whether fund managers can generate significant abnormal returns on the aggregate level through three different factor models. We select Carhart's (1997) four-factor model to estimate the funds' alpha, measuring abnormal performance. Second, we test whether the performance results from luck or skill through the bootstrap procedure of Kosowski et al. (2006) and test the null hypothesis of zero performance individually. Third, we use a similar bootstrap approach to Fama & French (2010) to jointly test the best and worst funds by the null hypothesis of zero performance, which allows us to evaluate if any of the best funds and their managers exhibit skill. Through this procedure, we account for possible correlation in the α that the factor model fails to capture. Ultimately, to present novelty to the subject, we implement the double bootstrap method of Harvey & Liu (2020). Their procedure allows us to estimate the Type II error rate and the Test power through the null hypothesis: a fraction of managers outperform or underperform. We follow their procedure to estimate cutoff values by controlling for a certain level of Type I error to characterize the Type II error rate.

Harvey & Liu (2020) and their recently developed performance evaluation method is the latest in a long line of methodologies with the same intent: to create a robust, definite and comprehensive statistical analysis towards mutual fund managers' actual performance and skill. The novelty of the methodology of Harvey & Liu (2020) lies in their assumptions on hypotheses. Instead of using a single hypothesis, they introduce the possibility of multiple alternative hypotheses in each test. Kosowski et al. (2006) only focus on Type I error rate, and their null hypothesis is that all funds have $\alpha = 0$. Harvey & Liu (2020) focus on both Type I and Type II error rates, and their null hypothesis is that a fraction of funds have $\alpha > 0$. By changing the null hypothesis, it is possible to make several tests with different assumptions upon the dataset. By executing the previously mentioned methods, we can test whether there are any outperforming or underperforming Norwegian mutual fund managers as a result of stock-picking skills or lack thereof.

Our findings suggest that we cannot identify any statistical significance in the positive alpha (0.035) on the aggregate level. When we estimate the standard parametric alpha for each fund, some Norwegian fund managers generate positive alphas, which implies that they beat their benchmark. However, when we test the null hypothesis of zero performance, we can distinguish skill from luck in the generated alphas. In the tests similar to Kosowski et al. (2006) and Fama & French (2010), we fail to reject the null hypothesis of zero performance for the funds that generate a positive alpha; we fail to identify skill. However, we reject the same null hypothesis for the funds that generate a negative alpha. The rejection implies that the worst performances are a result of the absence of skill.

By testing the null hypothesss: a fraction of managers outperform, in the double bootstrap approach by Harvey & Liu (2020), we reject the null hypothesis. We find no statistically significant evidence of skillful managers among the best performers. However, our estimated Test power of around 50% indicates that there might be skillful managers, but our test lacks the power to detect these. For the worst-performing funds, we fail to reject the null hypothesis of underperforming funds. Thus, we find statistically significant evidence that skill is absent among the worst-performing Norwegian mutual fund managers.

The rest of the thesis is organized as follows. Section 2 describes relevant literature, providing a brief literature review of previous papers on the subject. Section 3 contains the methodology of our thesis, describing the models and methods in detail. Section 4 presents our analyses' data, with a thorough presentation of our data sample, factors, and the Norwegian mutual fund market as a whole. Section 5 shows the empirical results of our findings and evaluations of the performance of mutual fund managers through the different statistical procedures and methods. Finally, section 6 contains remarks and conclusions.

2 Literature review

The literature on fund performance evaluation is highly influenced by the Capital Asset Pricing Model, CAPM. Through an expansion of the CAPM derived from the work of Sharpe (1964) and Lintner (1965), Jensen (1968) introduced Jensen's alpha, α. This approach aims to measure fund managers' performance by comparing actual returns and expected returns, conditional on the risk-free rate, systematical risk, and the actual return on the market portfolio in a specific time period. This model is commonly referred to as the single-factor model.

Using the single-factor model and the alpha as a measure of fund performance compared to the benchmark, Jensen (1968) analyzed US funds' historical return from 1945-1964. The assumption of Jensen (1968) was that investors would accept higher risk only if they are compensated with a higher expected return in the future. Therefore, the analysis incorporated the difference in returns of funds, regarding the Market factor and alpha of the fund, compared to a mean-variance efficient of an uninformed investor. The latter being the study's benchmark (Jensen, 1968). By using the market proxy of a passively managed fund, the analysis found that US funds, on average, were not able to beat their respective benchmarks sufficiently, considering the incurred management fees. Thus, the conclusion was that Jensen (1968) did not locate any significant positive alphas in his study. Hence, no outperforming funds were discovered using the single-factor model through multiple conventional testing.

Fama & French (1993) created the three-factor model to summarize and incorporate the cross-section of market stock returns in the mid-1990s. They identified that the cross-section average returns on US stocks displayed little relation to the market beta (β) of the regression in the CAPM model by Sharpe (1964) and Lintner (1965). Therefore, two additional factors, the Small-Minus-Big (SMB) and the High-Minus-Low (HML), were implemented to increase power in detecting managers' performance in mutual funds. Carhart (1997) argued, from the work of Grinblatt & Titman (1992) and Jegadeesh & Titman (1993), that the Momentum factor of stocks must be considered for the evaluation of mutual fund managers' performance, and thus introduced the four-factor model for asset pricing.

Gruber (1996) was the first to implement the multi-factor model in a study from 1985 to 1994. He argued that by looking at the average expense ratio, mutual fund managers could generate abnormal returns gross of expenses. The results suggested that mutual fund managers have superior stock-picking skills. However, the cost of these skills was too high. The question of whether mutual fund managers had sufficient stock-picking skills to cover their costs was further studied by Daniel, Grinblatt, Titman and Wermers (1997). They discovered that the mutual fund managers exhibit a certain level of stock-picking skill, identifying similar results to Gruber (1996), where managerial fees neutralized performance. Until this point, several studies differ in their results on whether there exist managers with stock-picking skills on an aggregate level. The results presented on the average level for mutual fund managers indicated that managers did not exhibit stock-picking skills to justify their management cost. However, the results do not imply that every mutual fund manager fails to beat their benchmark.

Kosowski et al. (2006) introduced a new approach to assess whether the abnormal return of mutual fund managers resulted from stock-picking skills or luck. Through their bootstrap method, they use an individual fund's bootstrapped p-value as a measurement to detect whether the identified abnormal return, positive or negative, was a product of skillful stock-picking or a result derived from lucky investments. Their results on the US fund market found that a minority of managers do, in fact, exhibit skill and produce returns to more than cover their cost, and keeping their positive α persisting over time. The methodology of Kosowski et al. (2006) will be discussed in detail later in the thesis. Fama & French (2010) further contributed to this bootstrap procedure with an alternative approach. They use a similar sample to Kosowski et al. (2006) and present the same results on the worst funds; they display a lack of stock-picking skill rather than an unlucky performance. However, they could not detect any skill among managers in contrast to Kosowski et al. (2006). Fama & French (2010) contributed to the joint test by Kosowski et al. (2006) to account for the correlation in the alpha estimates that the benchmark factor model did not capture. In addition, they suggested that the minimum required observations should be 8 months instead of Kosowski et al.'s (2006) 18 months to account for potential survivorship bias.

Harvey & Liu (2020) present a new approach to determine statistical significance in multiple testing of performance evaluation. By introducing the cutoff t-statistic as a measure of significance, Harvey & Liu (2020) control for a certain level of False Discovery Rate (FDR) to incorporate the possibility of luck in the sample and thus increase the threshold for significance. Their new procedure argues that if we cannot reject the null hypothesis, it may not necessarily be due to a lack of skill by managers. Instead, it may be a result of low power in the analysis performed. Harvey & Liu (2020) further discuss the method of cutoff t-statistics as a measurement for significance and expand their study by assessing the various error rates in multiple testing in performance evaluation. Creating a dynamic vector of funds being tested, p_0, introduced multiple testing with multiple hypotheses. They test for different numbers of alleged outperforming funds by using their method of cutoff t-statistic and introduce the assessment of Type I and Type II error rates in the sample. Harvey & Liu (2020) argue that the possibility of these error rates in performance evaluation may have a considerable economic impact on the investor. They introduce a tool, Oratio, to represent the trade-off between the error rates and to be used by investors to weigh the cost of these possible errors.

In the performance evaluation of the Norwegian Mutual Funds market, Sørensen (2011) employed the bootstrap method from Kosowski et al. (2006) with some modifications from Fama & French (2010). In addition, he implemented the factor models of Fama & French (1993) and Carhart (1997) onto the Norwegian mutual fund market. Using a sample free of survivorship bias, Sørensen analyzed the distinction between luck and skill in the Norwegian market. The results displayed weak tendencies towards skill in the cross-sectional distribution of alphas. However, he found clear signs of managers showing a lack of skill. Due to the small size of the Norwegian Fund Market compared to the US market, Sørensen (2011) focused on survivorship bias and implemented funds with a short lifetime. He argued that the overall Norwegian funds market would be biased if long life expectancy were put forth as an assumption for the sample with its low number of contestants. He further argues that a fund's short life span is explained by either administrative faults, unskilled management or insufficient management costs to cover their overall expenses. The findings also disclosed that there was little to no persistence regarding winners or losers in the market.

Previous research on Norwegian mutual fund performance portrays a certain amount of scarcity in the literature. However, over the recent years, there have been several master's theses on the subject, applying various procedures towards evaluating the performance of Norwegian mutual funds. Utseth & Sandvik (2015) performed an analysis of mutual fund returns, persistence and business cycle asymmetries in the Norwegian fund market by locating abnormal performance and whether such performance is considered skill or luck. The sample consists of 98 funds, with the lowest number of observations per fund being 17. Utseth & Sandvik (2015) implemented the bootstrap procedure of Kosowski et al. (2006). However, in their use of the factor model, they perform a consideration of both a conditional and unconditional use of the Carhart (1997) four-factor model but ultimately perform their analysis using the unconditional method. The unconditional four-factor model assumes constant exposure to the risk factors over time. Their results indicate no superior stock-picking skill in the Norwegian mutual fund market. However, they locate evidence towards, but ultimately reject, their null hypothesis of funds underperforming due to lack of skill. Further, they find evidence of persistence both in out- and underperforming funds up to a time window of 6 months. Ultimately, they identify evidence that fund managers' performance indicates some superiority to passive management in times of recession (Utseth & Sandvik, 2015).

Bråthe & Bjerke (2019) perform a similar study to Utseth & Sandvik (2015) on the performance in the Norwegian mutual fund market. The sample consists of funds' historical returns from 1987-2019. They keep the restrictions on the number of observations, 12 being the least number of observations of any fund in the sample. Bråthe & Bjerke (2019) also use the unconditional Carhart (1997) four-factor model in their analyses and the method of Kosowski et al. (2006) as their bootstrap approach. However, differing from the findings of Utseth & Sandvik (2015), they identify a clear indication of the lack of skill in active management of mutual funds. Moreover, their results on outperformance in funds coincide, providing evidence that no funds outperform due to skill, net of management fees.

3 Methodology

This section will describe the methodology of the thesis and the methods used in our analyses on the performance of Norwegian Mutual Funds. First, we will define the various benchmark models. Second, we will describe the bootstrap methods introduced by Kosowski et al. (2006), Fama & French (2010) and Harvey & Liu (2020) in detail, which we later perform in the Empirical Results.

3.1 Factor models for Performance of Norwegian Mutual Funds

In previous literature on the subject, there are several different approaches to the factor models of mutual fund performance. Jensen(1968) based his risk-adjusted measure, known as Jensen's Alpha, upon the Capital Asset Pricing Model (CAPM) by Sharpe (1964), Lintner (1965) and Mossin (1966). The single-factor model contains only one factor, MKT, and is given by

$$R_{i,t} = \alpha_i + \beta_i MKT_t + \epsilon_{i,t}, \tag{1}$$

where $R_{i,t}$ represent the return, α_i represent the funds return not explained by the Market factor, MKT represent the market risk premium and $\epsilon_{i,t}$ is the error term for a given fund, i at a given time, t. The β_i represents the coefficient value estimated through the OLS regression in the benchmark factor model. The Market factor is defined as the Market Risk Premium or the excess market return. In essence, MKT is computed as the benchmark index, here the OSEAX index, minus the risk-free rate. The choice of the benchmark index and the computation of the risk-free rate will be discussed in detail in Section 4 - Data.

Fama & French (1993) is arguably the most influential paper on performance evaluation, according to Sørensen (2011). Fama & French (1993) construct three different factors used in performance evaluation. First, they re-use the well-known factor of Market Risk Premium, the Market Factor (MKT), introduced by Jensen (1968). Second, they consider that small-capitalization stocks outperform large-capitalization stocks and thus construct the factor Small-Minus-Big (SMB). Third, they considered that stocks with a high book-to-market ratio often performed better than their counterpart and implemented the factor of High Minus Low (HML). The Fama & French three-factor model is given by

$$R_{i,t} = \alpha_i + \beta_i MKT_t + \beta_{2i} SMB_t + \beta_{3i} HML_t + \epsilon_{i,t}, \tag{2}$$

where α_i is the isolated performance parameter for a given fund, i, MKT is the Market factor, SMB is the Small-Minus-Big, HML is the High-Minus-Low and $\epsilon_{i,t}$ is the error term for a given fund, i at a given time, t. The β_i represent the coefficient value to be estimated through the OLS regression in the benchmark factor model.

In addition to the factors used in Fama & French's three-factor model, Carhart (1997) introduced the four-factor model. The new factor was estimated based on the findings of Jegadeesh and Titman (1993). By adding the Momentum factor (PR1YR), Carhart created the four-factor model containing the factors as parameters for the intercept. The Momentum factor focuses on the one-year performance anomaly by assuming stocks that previously have increased will continue to grow, and vice versa with decreasing stocks. Descriptive statistics for these factors are described and presented in detail under Section 4.4 Risk Factors. Carhart's four-factor model is given by

$$R_{i,t} = \alpha_i + \beta_{1i} MKT_t + \beta_{2i} SMB_t + \beta_{3i} HML_t + \beta_{4i} PR1YR_t + +\epsilon_{i,t}, \tag{3}$$

where the parameters remain as defined in Fama & French (1993) three-factor model, and PR1YR is the Momentum factor at time t.

All featured factor models contain the α_i as the pricing error. Thus, by definition, the α_i serves as the parameter which the factors fail to explain in computing the funds' excess return. In other words, after every factor in each model is taken into account, the remaining explanatory effect towards a funds return will be the α_i. Therefore, this term is the definition of excess return in our thesis, and we will focus on the alpha of each fund to assess its performance in the Norwegian Mutual Fund market.

Carhart's (1997) four-factor model will be the basis for our analyses, as we will present later in the thesis. The different factor models featured in this section will be discussed and considered in detail during Section 4 - Data. Then, we will discuss the different methods by their descriptive capabilities and statistical significance to further assess the use and choice of a benchmark proxy.

3.2 Bootstrap

This subsection presents the three bootstrap procedures performed in this thesis. The bootstrap methodology is essential to distinguish luck from skill in manager performance. First, we will describe the bootstrap method of Kosowski et al. (2006) to analyze statistical significance for all the individual mutual funds' performance. Second, we use Fama & French (2010) method to find statistical significance in a joint test for the best and worst-performing funds. Eventually, we will implement the method of Harvey & Liu (2020), where we find the cutoff t-statistics under the assumption that a fraction of managers have skills. This approach also allows us to estimate the Type II error and Test power.

Kosowski et al. (2006) present, among several arguments, the benefits of using the bootstrap approach when the sample size is small when there is a limited amount of observations. Our dataset of Norwegian mutual funds is quite limited, as presented in Appendix D. Both Kosowski et al. (2006), Fama & French (2010) and Harvey and Liu (2020) perform their analyses onto the US mutual fund market with a more comprehensive sample in terms of available data. In addition, some Norwegian mutual funds contain few observations due to a short life span. The limitations on data further display the necessity of a bootstrap procedure to conduct our hypotheses tests.

By repeating the same test of alphas several times, we retrieve the funds' alpha t-statistics. Some of the individual mutual funds display alpha values with nonnormally distributed returns. According to Kosowski et al. (2006), this nonnormality could arise for several reasons.

- Mutual fund managers tend to hold heavy positions in relatively few stocks or industries that contradict the central limit theorem where an equal-weighted portfolio of nonnormally distributed stocks will approach normality.
- Co-skewness between the market benchmark and the individual stock returns may occur. If the market benchmark returns are non-normal, then as a consequence, the individual stock returns may also be non-normal.
- Mutual fund managers tend to have dynamic strategies, changing their level of risk, in response to performance compared to similar portfolios or overall market portfolio changes.

Each of the irregularities presented could lead to non-normally distributed alphas, such that normality might be a poor approximation. Further, the cross-section has several conditions too. The correlation needs to be equal to zero in the residuals. The managers need similar risk levels to ensure no parameter estimator errors that undertake standard critical values of the normal distribution remain suitable for the cross-section. Considering all conditions as previously mentioned, the bootstrap procedure is the most suitable method to apply, and thus ensure that we do not have to rely on these parametric assumptions. The intention is that we instead estimate the statistical distribution of interest to infer in the multiple tests correctly.

In the first test by Kosowski et al. (2006), we evaluate both the estimated alpha and the alpha t-statistic. For the tests by Fama and Fench (2010) and Harvey & Liu (2020), we only consider the alpha t-statistic. According to Busse, Goyal & Wahal (2010), Fama & French (2010) and Kosowski et al. (2006), there are statistical reasons to evaluate the alpha t-statistic rather than the alpha estimates. These reasons are due to alpha estimation's precision, where the alpha will vary across the funds for two reasons—the length of mutual funds' return history and the degree of diversification. The alpha t-statistic (α_t) describes the statistical significance of the estimated alphas.

3.2.1 Kosowski et al. (2006)

We use a bootstrap method similar to Kosowski et al. (2006) and operate under the null hypothesis of no true performance in any individual fund. The bootstrap procedure estimates a simulated distribution of each fund's alpha. From the distribution of alphas, we can generate a bootstrapped p-value which allows us to distinguish between luck and skill in the performance of funds. Essentially we want to test the null hypothesis of zero true performance: $\alpha_i = 0$ or $\hat{t}_{\hat{\alpha}_i} = 0$ for each fund i.

The bootstrap method estimates the distribution by resampling the monthly returns multiple times with replacement from the original sample and compute the t-statistic of each resample. In the bootstrap procedure, we use Carhart (1997) four-factor model, which is given by

$$R_{i,t} = \hat{\alpha}_i + \hat{\beta}_{1i}MKT_t + \hat{\beta}_{2i}SMB_t + \hat{\beta}_{3i}HML_t + \hat{\beta}_{4i}PR1YR_t + \hat{\epsilon}_{i,t}, \tag{4}$$

to compute ordinary least squares for estimation of alphas, factor loadings and the residuals for each fund i.

Before the bootstrap procedure we need to retrieve the following coefficients for each fund i: ($\hat{\alpha}_i$,$\hat{\beta}_{1,i}$,$\hat{\beta}_{2,i}$,$\hat{\beta}_{3,i}$,$\hat{\beta}_{4,i}$) and the t-statistic of alpha ($\hat{t}_{\hat{\alpha}}$). We also save the time series of estimated residuals: $\hat{\epsilon}_{i,t}$, where $t = T_{i0}, ..., T_{i1}$ and T_{i0} and T_{i1} is the dates for the first and last available monthly return for each fund i.

We draw a random sample with replacements from the funds' residuals saved from the OLS regression. Thus creating a pseudo-time series of resampled residuals, $\hat{\epsilon}^b_{i,t_\epsilon}$ where b represent the bootstrap number ($b = 1, \ldots, 1000$). Each of the time indices $s^b_{T_i0}, ..., s^b_{T_i1}$, are drawn randomly from $T_{i0}, ..., T_{i1}$, where the original sample of each fund are reorganized. The estimated factor returns remain unchanged and have the same chronological ordering in the constructed dataset. The next procedure is to impose the null hypothesis of zero true performance in the time series of pseudo-monthly returns, $\alpha_i = 0$, by

$$R^b_{i,t} = 0 + \hat{\beta}_{1i} MKT_t + \hat{\beta}_{2i} SMB_t + \hat{\beta}_{3i} HML_t + \hat{\beta}_{4i} PR1YR_t + \hat{\epsilon}^b_{i,t_\epsilon}, \tag{5}$$

to be able to test if the actual alpha or alpha t-statistic is equal to zero. The returns are regressed for a given bootstrap sample, b, on the factors from Carhart (1997) four-factor model. A positive or negative alpha and t-statistic of alpha may be presented if an abnormally high or low number of residuals are drawn. We repeat this for all the funds $i = 1, \ldots, 107$, which results in a bootstrapped alpha for each fund. This procedure is then repeated 1000 times. Next, we arrange the alpha values ($\hat{\alpha}^b_i, i = 1, 2, \ldots, 107$) and the t-statistic ($\hat{t}^b_{\hat{\alpha}i}, i = 1, 2, \ldots, 107$) from highest to lowest value for each bootstrap, b, achieving a result purely from sampling variation. We use the p-value to see if the bootstrap iteration generate more extreme positive values of estimated alpha $\hat{\alpha}^b_i$ or t-statistic of alpha $\hat{t}^b_{\hat{\alpha}i}$, compared to those observed in the actual data. We reject the null hypothesis implying that the sampling variation is the source of high alphas and that the result is not the result of luck. We calculate the p-value for the funds that generate a positive alpha by

$$P - Value_i = \frac{\sum_{i=1}^{R} \alpha_{\hat{i}} > \widehat{\alpha}_i}{R}, \tag{6}$$

where i represent each fund individually and we summarize the number of observations where α_i is higher than the estimated alpha, $(\widehat{\alpha})$, and divide it by the number of simulations R. When we calculate the p-value for the funds that generate a negative alpha we use $\alpha_{\hat{i}} < \widehat{\alpha}_i$ in the same equation.

3.2.2 Fama & French (2010)

In this section, we will focus on the two outer tails of the ranked bootstrap distribution. We follow a similar approach to Fama & French (2010) to test the null hypothesis of no true performance for all the funds or the alternative hypothesis that the average alpha for the top or bottom performers is not equal to zero. Fama & French (2010) suggest a method to sample funds and the explanatory returns jointly. While Kosowski et al. (2006) perform independent simulations for each fund, Fama & French (2010) perform a joint sampling of fund returns and the explanatory returns. Fama & French (2010) argue that the method of Kosowski et al. (2006) does not account for the correlation of α-estimates by the potential correlation that arises due to the benchmark model not capturing all common variation in the funds' return. Thus, the method would result in missing effects of correlated movement in the volatility of the factor returns and residuals. However, these accusations are wrong; Kosowski et al. (2006) examine these concerns in their paper. However, our approach is more comparable with Fama & French (2010) because the minimum requirement of observations is more similar than Kosowski et al. (2006).

The methodology is quite similar to the individual bootstrap procedure. However, they differ in the number of alpha t-statistic for each fund to compute the p-value. In essence, Fama & French (2010) use the ranked t-statistic and bootstrap results to find the average of a percentile of the highest or lowest t-statistics to compute the p-value. Further, the p-value allows us to measure whether the actual performance is perceived as extreme compared to the performance in the bootstrapped results. For example, if a fraction below 5% of the simulations produces alpha t-statistics higher than the actual estimates, the result would indicate that a fraction of managers have skills. The formula for p-value remains as presented above, except that we test the mean of the best and worst generated alphas and its distribution of statistics under the null hypothesis. We measure the number of observations where the mean of the groups generated alphas are above the estimated mean of each simulation and divide it by the number of simulations to estimate the p-value for the best funds. Further, we count the observations when our test statistic is below the distribution of this statistic under the null hypothesis to estimate the p-value for the worst funds.

3.3 Harvey & Liu (2020) bootstrap procedures

So far, we have discussed the multiple testing adjustment with single hypotheses tests of Kosowski et al. (2006) and Fama & French (2010). They evaluate the performance of mutual funds and test the null hypothesis of zero alpha for all funds, $\alpha = 0$, to distinguish luck from skill. However, neither of these procedures assess the probability of falsely declaring that no manager has skill when some funds display a positive alpha, the Type II error.

First, we introduce the methodology of cutoff t-statistic by Harvey & Liu (2020). By estimating a statistical threshold to control for luck, we focus on the Type I error rate, the false discovery rate (FDR). Second, we consider the methodology by Harvey & Liu (2020), incorporating Type II error rate estimation in multiple testing, allowing us to estimate the Test power. We conduct the double bootstrap approach and operate under the null hypothesis that a fraction of managers outperforms or underperform.

3.3.1 Cutoff t-statistic

In the first test, we follow the approach of Harvey & Liu (2020) to test the null hypothesis of no true performance for all the funds against the multiple alternative hypotheses that a specific amount of managers is outperforming or underperforming. Table 1 illustrate the various alternative hypotheses we test.

Table 1: Hypotheses by Harvey & Liu (2020)

Null (H_0)	Alternative (H_A)	
$H_0 : \alpha_i = 0$	H_A^1:	$\alpha_1 > 0$
	H_A^2:	$\alpha_1 > 0$ $\alpha_2 > 0$
	H_A^N:	$\alpha_1 > 0$ $\alpha_2 > 0$ $\alpha_N > 0$

In Table 1, i represents the alphas of each of the 107 mutual funds. N represents the number of alternative hypotheses we want to test. The interpretation of the table is that we test the null hypothesis of $\alpha = 0$ for all the funds against the alternative hypothesis, i.e., H_A^2 that the two best funds have positive alphas. Then, we use the same alternative hypothesis to test whether the worst performing fund has $\alpha < 0$ further, whether the two worst funds have $\alpha < 0$, we repeat for N alternative hypotheses.

First, we perturb the original data X_0 by bootstrapping the time periods through 1.000 simulations to create an alternative panel of returns X_i. Further, we rank the funds in a descending order based on their t-statistics, rendering the matrix manageable for us to test multiple alternative hypotheses of more than one outperforming fund under the null hypothesis: $\alpha = 0$. In the first test, we operate under the prior that zero funds outperform or underperform, given by $p_0 = 0$. When $p_0 = 0$, the approach is similar to that used in Kosowski et al. (2006) and Fama & French (2010) by creating the ”pseudo” sample of fund returns. As described above, the sample of returns is applied to each hypothetical level of alternative hypotheses, differing on the prior on p_0, to estimate the cutoff t-statistics between 1.5 to 4.0 with 0.1 increments for each alternative hypothesis. We can perform the same procedure for negatively performing funds to find the cutoff t-statistic between -4.0 to -1.5 with 0.1 increments. To estimate cutoff t-statistics, we control for the False Discovery Rate (FDR). The FDR is computed as

$$FDR(t) = \frac{\sum_{j=1}^{J} h_j}{J}, \tag{7}$$

where t represents the cutoffs from 1.5 to 4.0 with 0.1 increments, j represents each simulation, J represents the total number of simulations, and h is the number of times our bootstrap results are higher than the threshold. These observations are then summarized and divided by the total number of simulations, creating the false discovery rate. We use the number of simulations to fulfill h_j divided by the total number of simulations to generate the False Discovery Rate per hypothesis. The FDR displays the ratio of false discoveries within the bootstrapped sample dependant on the alternative hypotheses. A higher number of simulations with the rejection of H_0 displays a higher potential false discovery rate.

Using the FDR and the t-statistics, we can compare the 1.000 simulations of all 107 funds to identify how many funds receive a t-statistic alpha higher than the cutoff. By identifying whether or not one or more funds outperform, we can compare these results with the FDR Matrix of each fund as displayed in Appendix F. The cutoff t-statistic is determined by controlling for a specific level of FDR. In our analysis, we control for 5 percent of the Type 1 error rate. The result will be 0.05 FDR at the cutoff t-statistic distribution for all hypotheses H_A^N.

3.3.2 Double bootstrap - error rates

Harvey & Liu (2020) suggests a double bootstrap procedure which allow us to estimate both error rates, Type I and Type II. The double bootstrap procedure assumes $\alpha \neq 0$ on a prior of funds, p_0. The p_0 represents a researcher's or an investor's prior on the number of out- or underperforming funds for a given sample. We set a prior on p_0 and bootstrap the data a second time. The p_0 number of funds retain their respective alphas, and makes it possible to assess the rate of False Positives and False Negatives in the sample. The double bootstrap approach estimates the rate at which these errors occur in performance evaluation of mutual funds. The hypotheses in this method are given by

$$H_0 : \alpha_{p0} \neq 0, \tag{8}$$

$$H_A : \alpha_{p0} = 0, \tag{9}$$

where p_0 represents the prior on outperforming or underperforming funds, and where the H_0 is testing for significant alphas based on our prior beliefs. In essence, p_0 defines the number of funds retaining their alpha in the bootstrap, and as a result, we assess whether or not p_0 of funds are outperforming or not.

Table 2 displays the overall testing outcomes of a double bootstrap procedure. $FP^{i,j}$ is an observation of a false positive identified through i, first bootstrap and j, second bootstrap. $TP^{i,j}$ is the identification of a true positive outcome from both rounds of bootstraps, $TN^{i,j}$ is the identification of a true negative outcome and $FN^{i,j}$ is a false negative outcome estimated from both bootstraps. We use the double bootstrap procedure as we are interested in both Type I and Type II error rate of our testing procedures. In addition, the double bootstrap simplifies the use of alternative hypotheses by grouping similar alternative hypotheses. The alternative hypotheses, which correspond to a fraction of p_0 to be true, are grouped under a single H_A associated with the P_0. Finally, the second-stage bootstrap will allow us to estimate both Type I and Type II errors nonparametrically.

Table 2: A contingency table for testing outcomes

Decision	Null (H_0)	Alternative (H_1)
Reject	False positive (Type 1 error) $(FP^{i,j})$	True Positive $(TP^{i,j})$
Accept	True negative $(TN^{i,j})$	False negative (Type 2 error) $(FN^{i,j})$

Harvey & Liu (2020) suggest the following steps to perform multiple testing with error rates being assessed. The following itemized paragraph is a direct quote from Harvey & Liu (2020), p. 2513:

- ”Step I. Bootstrap the time periods and let the bootstrapped panel of returns be X_i. For X_i, obtain the corresponding $1 \times N$ vector of t-statistics t_i.
- Step II. Rank the components in t_i. For the top p_0 of strategies in t_i, find the corresponding strategies in the original data X_0. Adjust these strategies so they have the same means as those for the top p_0 of strategies ranked by t_i in X_i. Denote the data matrix for the adjusted strategies by $X_{0,1}^{(i)}$. For the remaining strategies in X_0, adjust them so they have zero mean in-sample (denote the corresponding data matrix by $X_{0,0}^{(i)}$). Arrange $X_{0,1}^{(i)}$ and $X_{0,0}^{(i)}$ into a new data matrix Y_i.
- Step III. Bootstrap the time periods J times. For each bootstrapped sample, calculate the realized error rates (or odds ratio) for Y_i, denote by $f_{i,j}$ (f stands for a generic error rate that is a function of the testing outcomes).
- Step IV. Repeat Steps I to III I times. Calculate the final bootstrapped error rate as”

$$\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} f_{i,j}, \tag{10}$$

where I is the number of simulations in the first bootstrap, J is the number of simulations in the second bootstrap, i represents the observed values from the first bootstrap, and j represent the observed values from the second bootstrap. Finally, f represents the general parameter for the different error rates.

We will estimate the error rates that may occur in multiple testing. Type I implies the selection, with the assumption of skillfulness, of a manager or a strategy that turns out to be unskilled. Such an error may have variations in economic implications, where a manager may slightly underperform or, worse, have a significant negative return. Thus, the Type I error rate is often referred to as a false positive. Several methods take the Type I error rate into account, assuming that this error is the most expensive, economically speaking. Type II error rate is referred to as a false negative. In essence, the rejection of a manager due to assumed under-performance when there is not. Harvey & Liu (2020) found that current performance evaluation methods ignore the Type II error. However, the economic implications of the two error rates may differ considerably. For example, a lost opportunity may be more costly than accepting a failing strategy, depending upon the scenario at hand. As of current methods, investors have no current tool to evaluate the different possible error rates. The lack of an evaluating framework of these error rates is what Harvey & Liu (2020) present a solution to. The Oratio represents the trade-off between the two error rates. It provides the rate of false discoveries per miss (Harvey Liu, 2020). Oratio takes the value and size of p_0 into account, thus providing a dynamic representation of the error rates. For example, using the Oratio as a measure of risk, an investor could say that Type I error rate is ten times more costly than Type II and thus desire an Oratio of 0.1 to evaluate the cutoff t-statistic for different values of p_0.

We can estimate the error rates for various null hypotheses through the double bootstrap procedure where a fraction, p_0, is assumed outperforming or underperforming. First, we use the realized FDR (RFDR), which is defined as

$$\text{RFDR}^{i,j} = \begin{cases} \frac{FP^{i,j}}{FP^{i,j}+TP^{i,j}}, & \text{if } FP^{i,j}+TP^{i,j} > 0 \\ 0, & \text{if } FP^{i,j}+TP^{i,j} = 0, \end{cases} \tag{11}$$

where i and j represent the first (I) and second (J) bootstrap simulations, respectively, FP is the estimation of false positives, and TP is the number of true positives. RFDR is the realized false discovery rate. Further, we take the average across the perturbed data i, and each time we generate j bootstrapped random samples. We can find the Type I error rate, which is the possibility of selecting a manager that turns out to be unskilled and compute

$$\text{Type I} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} RFDR^{i,j}, \tag{12}$$

where Type I represent the Type I error rate, I, J, i, j and RFDR are defined as previously mentioned.

The realized rate of misses (RMISS), also sometimes referred to as the false omission rate or false non-discovery rate, is defined by

$$\text{RMISS}^{i,j} = \begin{cases} \frac{FN^{i,j}}{FN^{i,j}+TN^{i,j}}, & \text{if } FN^{i,j} + TN^{i,j} > 0 \\ 0, & \text{if } FN^{i,j} + TN^{i,j} = 0, \end{cases} \tag{13}$$

where RMISS is the realized rate of misses, FN is the false negative observations, TN is the true negative observations and i and j remain as previously defined. RMISS is further used to find the Type II error rate, which is the possibility of not selecting or missing a manager that turns out to be unskilled but was not, given by

$$\text{Type II} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} RMISS^{i,j}, \tag{14}$$

where each variable remain as previously defined in the previous equations.

However, the Type II error rate holds different definitions and interpretations. The Type II error rate as defined by Harvey & Liu (2020) deviates from the traditional view of Type II error rate, where traditional error rate provides the possibility of estimating Test power as Test power = $1-$Type II. The Type II error rate from Harvey & Liu (2020) does not provide this possibility. To evaluate the Test power of our analysis, we will estimate the traditional, standard interpretation of Type II error rate, hereafter defined as S-Type II. We must retain both Type II error rates in our analyses, as Type II is the rate used in the framework of Harvey & Liu (2020), and we use the S-Type II for the estimation of our analysis' Test power, computed as

$$\text{S-RMISS}^{i,j} = \begin{cases} \frac{FN^{i,j}}{FN^{i,j}+TP^{i,j}}, & \text{if } FN^{i,j} + TP^{i,j} > 0 \\ 0, & \text{if } FN^{i,j} + TP^{i,j} = 0, \end{cases} \tag{15}$$

$$\text{S-Type II} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} S - RMISS^{i,j}, \tag{16}$$

where the computations differ from Type II by using TP, true positives, in the denominator, the computation of S-Type II will be used in the empirical analysis for Test power in the section of Error rates.

With both Type I and Type II error rates from Harvey & Liu (2020), we are able to define the realized ratio of false discovery to misses, RRATIO as

$$\text{RRATIO}^{i,j} = \begin{cases} \frac{FP^{i,j}}{FN^{i,j}}, & \text{if } FN^{i,j} > 0 \\ 0, & \text{if } FN^{i,j} = 0, \end{cases} \tag{17}$$

where RRATIO is the rate of false discovery to misses, FN is the observed false negative, FP is the observed false positive and i and j remain as previously defined. We use the RRATIO to find the Oratio, the odds ratio between false discoveries and misses, given by

$$\text{Oratio} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} RRATIO^{i,j}, \tag{18}$$

where Oratio represent the odds ratio trade-off between Type I and Type II error rate.

The Type II error rate is the rejection of a mutual fund manager under the assumption of outperforming or underperforming managers when the opposite is true. In effect, the Type II error rate tells us a lot about the robustness of a test. The higher the rate of Type II, the higher the possibility of skill among some managers. The idea is that, instead of controlling the error rates for all circumstances, the analyses should incorporate the prior beliefs.

4 Data

This section presents the data basis for our computations and analyses on the performance evaluation of Norwegian mutual funds. We explain the different parameters in the thesis and describe the dataset used for the Empirical Analysis in Section 5. First, we present the sample mutual funds. Second, we offer the basis for the interest rate, the basis for the selected benchmark, and the four risk factors in Carhart's (1997) four-factor model. Ultimately, we discuss potential biases which may occur in the performance evaluation of mutual funds. The dataset of funds' monthly returns, risk factors and interest rates are collected from the database of Ødegaard (2019).

4.1 Sample mutual funds

The data basis is comprised of Norwegian Mutual funds' financial data from 1987 to 2019. We have included both surviving and non-surviving funds, but with a minimum of 12 months of observations. In addition, only funds with a minimum of 80% domestic equities are included in the sample. Only actively managed funds are included, where passive funds such as index funds and similar products are omitted from the sample. The dataset contains 107 Norwegian Mutual funds that met the requirements within the time period stated. The starting point of observations is set at 1987 due to the highly unstable interest markets before this time. Due to high volatility and inconsistency in the interest rates, our empirical analysis would be heavily affected by these instabilities, rendering the analysis less robust. In addition, NIBOR, which is the Norwegian Interbank Offered Rate, is available historically from this year. Given our focus on the methodology of the analyses of Norwegian mutual funds, we find using a single interest rate proxy through the whole dataset complementary to a robust result. We aim to locate error rates in previous analyses of Norwegian mutual funds. We have an obvious limitation set only to include the Norwegian mutual Funds data.

Daily Net Asset Value from each fund is retrieved from the TILTON database. This data is used to calculate the net monthly returns of each fund, providing 12 observations per fund each year. Net Asset Value (NAV) is a specific fund's market value per share at a given time. Using NAV, the monthly return of each fund is calculated by the last observed value of stock for each month. When adjusting NAV by subtracting dividends and management fees, the net return of each fund will be provided without the fund-specific incurred costs, creating a comparative table of returns. Calculated by

$$r_t = \frac{NAV_t - NAV_{t-1}}{NAV_{t-1}}, \tag{19}$$

where monthly return (r) from time t_0 to t_1 is given by the change in adjusted NAV from the current month divided by adjusted NAV from the previous month subtracted by 1. Thus, adverse changes in NAV will result in a loss, whereas a positive return will increase a fund's market value per share, providing the fund's net result for the particular month.

From Appendix D, Mutual Fund Descriptive Statistics, the number of observations per fund varies greatly. Therefore, even though the life span and other fund statistics are drastically different, all funds are weighted equally to ensure comparability of performance regardless of the number of observations.

4.1.1 Historically on the Norwegian fund market

The historical development of the Norwegian fund market is presented in Figure 1. Quantitative data on assets and customer relationships of Norwegian funds are available from the year 2003, Verdipapirfondenes Forening (2021). As Norwegian mutual funds are open-ended equity funds, the data represent a market without accessibility limitations in trading for the individual investor. The historical data is divided into private and institutional sectors, where the private sector contains data from individual investors and legal individuals. In contrast, the institutional sector is mainly counties, companies and other administrative entities. ”Total assets managed” are the total assets managed by a fund at the end of a given year, whereas customer relationships are the number of individual stakeholders in each fund. One investor may have several customer relationships, as one individual investor may hold shares in more than one fund at a time, providing one relationship per ownership of the fund.

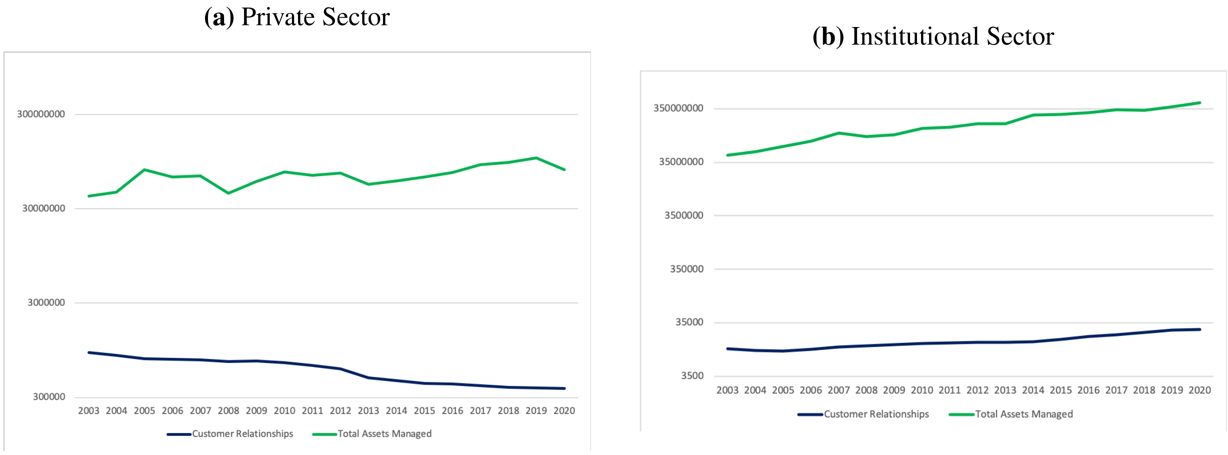
The data basis is comprised of Norwegian fund products only. Therefore, we have included Norwegian funds exclusively, hereby mutual funds, interest funds, hedge funds and bond funds as the chosen parameters. We aim to show the overall growth and decline in the fund market for private and institutional investors and highlight the changes in total assets managed and customer relationships at the end of every year from 2003 to 2020. The graphs are displayed logarithmically to enclosure both total assets managed and customer relationships in the same frame for comparative reasons.

There has been an overall decrease in customer relationships in the private sector from 2003 to 2020, with a negative development from 901,796 to 376,347. However, as this is Norwegian funds exclusively, this could mean that there has been a change in the investment pattern to a more international investment focus among Norwegian fund customers. It could also be explained by decreased diversity, as customers could tend to invest in fewer funds than before. However, there is no information on diversity available due to privacy legislation, and thus remain speculations. As the number of customer relationships decreases, total assets managed increases steadily over time, except for 2007 and 2008 where the world’s financial markets suffered from the financial crisis. Therefore, increased total assets managed and decreased customer relationships represent increased individual positions, implying that fewer investors invest a higher amount in Norwegian funds.

The institutional sector displays an increase in both assets managed and customer relationships. Total assets managed have experienced a noticeable increase from 47,927,913 in 2003 to 452,000,468 in 2020. This development grows steadily throughout the time period and depicts a growing market for institutional investors and their interest in the Norwegian fund market. Customer relationships have also seen a steady increase, resulting in a market in high demand.

Figure 1: Sector development 2003-2020

Figure 1 shows the development of customer relationship and total assets managed in the Norwegian mutual fund market from 2003-2020. The graphs are displayed logarithmically to ensure they are comparable, given the large differences in volume between customer relationships and total assets managed. Figure a) show the development of the private sector in Norwegian mutual funds trading, while Figure b) displays the development in the institutional sector. The private sector is defined as a private individual making trades in his or her legal name using the social security number as the identifier. The institutional sector consist of companies and businesses investing in Norwegian funds through their organizational number, where the company legally owns the papers on the fund. Institutional actors will also include municipal administrations, pension saving in businesses, counties and similar. Norwegian mutual funds are in the analysis defined through the categories predefined by VFF(2021) and further selected: Norwegian mutual funds, Norwegian money market funds, Norwegian combination funds, Norwegian hedge funds, and Norwegian fixed income funds.



(a) Private Sector

(b) Institutional Sector

Table 3 reports the market development in descriptive statistics, including net fund acquisition and assets per customer. Net fund acquisition is the acquisition amount of a fund subtracted redemption from one year to the next. Positive net fund acquisition is a growth in the invested amount in Norwegian mutual funds overall. Assets per customer is the relative ratio of total assets managed and customer relationships. As the number of customers decreases and total assets managed increases, assets per customer increase over time, pointing towards managing increasingly more prominent positions from the individual customer. For institutional customers, the tendencies are the same as in the private sector, increasing assets per customer. However, institutional customers report a more steadily positive net acquisition. Due to institutional customers may not being as sensitive to market movement, they increase net fund acquisition over time, showing a more robust market for growth.

Table 3: Private and institutional customers - Norwegian fund instruments only

The table presents descriptive statistics from the VFF database (2021) for the Norwegian mutual fund market by selecting Norwegian money market funds, Norwegian combination funds, Norwegian hedge funds, and Norwegian fixed income funds. The table displays yearly key information regarding the overall fund market for private actors. The sample contains a larger variety of funds to display the overall growth in the market of funds of Norwegian origin. The second part of the table displays the institutional sector, showing key information regarding trading Norwegian mutual and other funds for institutional investors, companies, municipalities, and counties. International funds are omitted from the sample to create a Norwegian product focus.

Year	Customer relationships	Total assets managed	Assets per customer	Net fund acquisition
Private				
2003	901,796	40,888,678	45	762,152
2004	833,884	44,794,097	54	-972,475
2005	777,259	77,434,560	100	24,516,594
2006	763,655	64,764,823	85	-18,118,968
2007	754,277	67,098,959	89	59,866
2008	724,086	43,897,919	61	-9,739,150
2009	735,225	58,174,630	79	4,166,038
2010	701,601	73,783,700	105	560,663
2011	653,494	67,455,912	103	153,801
2012	607,521	71,706,382	118	2,855,966
2013	483,217	54,424,178	113	9,900
2014	449,971	58,878,240	131	1,997,753
2015	423,087	64,962,374	154	1,849,955
2020	416,528	72,701,752	175	3,412,706
2017	398,763	87,965,536	221	10,652,785
2018	385,468	93,096,105	242	-2,661,965
2019	376,820	102,907,202	273	3,328,981
2020	376,347	77,328,966	205	-385,108
Institutional				
2003	11,290	47,927,913	4,245	6,926,678
2004	10,593	55,573,515	5,246	7,509,428
2005	10,431	68,394,241	6,557	8,603,788
2006	11,047	87,327,160	7,905	14,822,526
2007	3,959	28,656,363	7,238	158,285
2008	12,807	106,699,589	8,331	-3,528,530
2009	13,677	113,190,251	8,276	11,265,625
2010	14,276	150,745,868	10,559	15,188,860
2011	14,582	157,078,411	10,772	13,314,336
2012	15,043	184,983,408	12,297	19,390,427
2013	15,226	184,837,638	12,140	-9,307,861
2014	15,586	266,402,397	17,092	72,700,951
2015	16,962	273,654,221	16,133	12,180,953
2020	19,250	296,147,782	15,384	11,056,879
2017	20,899	335,221,802	16,040	22,811,577
2018	22,885	324,206,231	14,167	2,853,590
2019	25,207	377,305,828	14,968	33,373,969
2020	26,246	452,000,468	17,222	20,057,986

4.2 Risk-free rate

To calculate the excess return of each fund, we have to account for the risk-free rate in the market. Ødegaard (2011) argues that NIBOR, Norwegian Interbank Offered Rate, serves as a suitable proxy for risk-free rates in calculating the excess return of Norwegian mutual funds. NIBOR serves as the guiding rate used for unsecured money lending between banks. The historical development of NIBOR from 1987 to 2019 is displayed in Appendix A. As previously mentioned, the data period of NIBOR will range from 1987 to 2019 due to ”chaotic” rates in the period before 1987, such as a Norwegian currency crisis which would create irregularities in our analyses. By using monthly NIBOR interest rates, the risk-free rate to be deducted is calculated as

$$r_f = (1 + NIBOR)^{\frac{1}{12}} - 1, \tag{20}$$

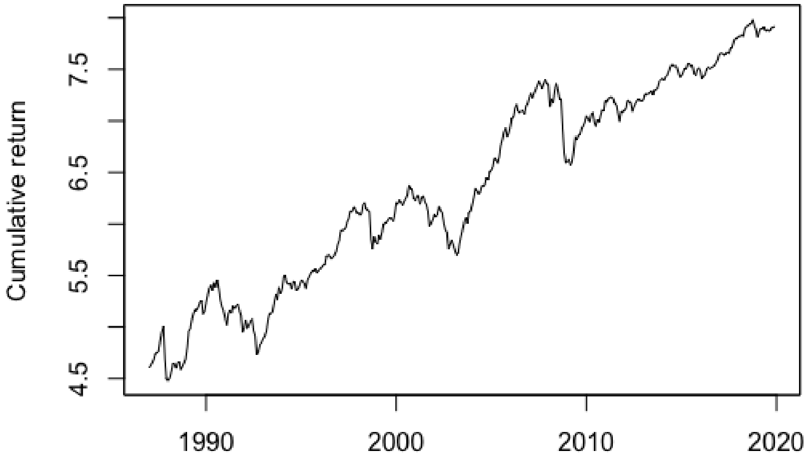
where r_f represent the risk-free rate and NIBOR represent the Norwegian Interbank Offered Rate at a given time.

4.3 Benchmark index

In assessing mutual fund performance, the excess return of the funds must be compared to a benchmark. The Oslo Stock Exchange provides various benchmarks. The Oslo Stock Exchange Mutual Fund Index (OSEFX) mirrors the movement of the stocks on the Oslo Stock Exchange following the UCITS standards. These standards include rules for diversification and limitations on the weight of the most extensive stocks on the exchange. The UCITS standard is part of an EU directive that aims to standardize and harmonize the management and sales of mutual funds. Therefore, the OSEFX benchmark would be a suitable benchmark in our analyses. However, the benchmark is relatively new, first introduced in 1995. For analytical purposes, our computations and comparisons require one benchmark to be used throughout the whole data period, and OSEFX is thus unsuitable. The Oslo Stock Exchange All-Share Index (OSEAX) serves our purpose and contains every stock listed on Oslo Stock Exchange adjusted for dividends and daily corrections by corporate actions. In our analyses, we will be using OSEAX as the Market proxy with a weighted portfolio reflecting close to all shares on the Oslo Stock Exchange, except for the least traded. Figure 2 illustrates the cumulative return of OSEAX for the entire time period we investigate.

Figure 2: Cumulative return on OSEAX

This figure plots the cumulative return of OSEAX from 1987 to 2019. It illustrates the benchmark development in a logarithmic scale.



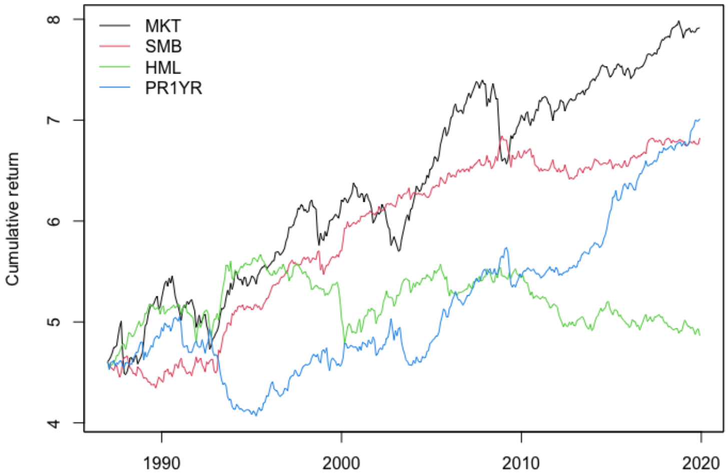
4.4 Risk factors

The risk factors in the Carhart (1997) four-factor model are collected from empirical data on Oslo Stock Exchange in Ødegaard's (2019) database. The factors of the model were described previously in the methodology Section 3.1. We have retrieved and applied the Market factor (MKT) and the Small-Minus-Big (SMB), that companies with a low average market value have a different return than companies with high market value. We also apply the value factor, High-Minus-Low (HML), the yield of companies with high book to marked ratio subtracted for the return of companies with low book to market ratio. Finally, the Momentum factor (PR1YR) considers that stocks with changing values will continue to have a changing value in the future. These factors enable us to calculate Ordinary Least Squares (OLS regression) using time series data from different factors to estimate the Carhart (1997) four-factor model.

Figure 3 illustrate the factor returns from 1987 to 2019 with cumulative returns. By assessing the cumulative return for the factors over the entire period from 1987 to 2019, the factor that produces the highest accumulated return is the Market factor (MKT). MKT is also the most volatile, delivering the highest standard deviation throughout the period. The HML reports the lowest cumulative returns from approximately 2010 and decreasing returns from 2000, indicating

Figure 3: Cumulative return on factors

This figure plots the cumulative return from Carhart's (1997) factors from 1987 to 2019. This illustrates the impact of the Market factor, Small-Minus-Big, High-Minus-Low and Momentum factor in a logarithmic scale.



that companies with a high book to market value produce lower yields than the companies with a low book to market value. The SMB and the PR1YR factors display stable positive growth from the mid-1990s, where PR1YR eventually surpassed the SMB factor.

Table 4 reports various statistics for the factors in the entire sample period, divided into three different time periods at 11 years each. In Panel A, the mean factor returns are presented. All factors except the HML generate exclusively positive mean returns. The negative return from 2009-2019 (-4,764) in the HML factor implies, as stated in Figure 3 that companies with low book to market had higher yield than companies with high book to market value. The SMB factor generates positive mean returns in every period indicating that small firms generate higher yields than big firms. There is less of a difference from 2009 to 2019. The Momentum factor (PR1YR) produced the highest generated mean return in every period and contained the highest generated mean return (13,579) in 2009-2019. The high generated mean return in the Momentum factor can be described by the high growth from the financial crisis in 2008 and supplemented by easier access to trading platforms, which may have created synergies that affect the Momentum factor.

Table 4: Descriptive statistics of factors

The table below illustrates summary statistics for the Carhart (1997) four-factor model in the full sample period and three different periods: 1987-2019, 1987-1997, 1998-2008 and 2009-2019. The time periods consist of 11 years and start from January to December. Panel A reports the annualized mean return, Panel B reports the standard deviation, Panel C reports the maximum and minimum to each factor in percent, and Panel D reports the correlation between all the factors.

	MKT	SMB	HML	PR1YR
Panel A: Mean factor return				
Jan 1987-Dec2019	7.094	7.853	2.079	8.674
Jan 1987-Dec1997	7.263	10.418	9.790	1.272
Jan 1998-Dec2008	2.428	11.874	1.213	11.170
Jan 2009-Dec2019	11.591	1.268	-4.764	13.579
Panel B: Standard deviation				
Jan 1987-Dec2019	20.736	14.159	16.177	16.305
Jan 1987-Dec1997	22.741	15.755	18.199	17.234
Jan 1998-Dec2008	24.178	14.239	17.798	17.559
Jan 2009-Dec2019	13.829	12.165	11.494	13.755
Panel C: Max(min) factor returns				
Jan 1987-Dec2019	16.5(-28.7)	22.1(-17.1)	14.7(-16.7)	15.4(-16.8)
Jan 1987-Dec1997	16.5(-28.7)	22.1(-10.5)	14.7(-15.7)	13.5(-16.8)
Jan 1998-Dec2008	11.8(-24.6)	13.3(-17.1)	9.3(-16.7)	15.4(-14.2)
Jan 2009-Dec2019	14.9(-9.0)	12.6(-11.0)	6.9(-7.4)	12.1(-16.1)
Panel D: Factor correlation matrix				
MKT	1	-0.448	0.051	-0.157
SMB	-0.448	1	-0.138	0.105
HML	0.051	-0.138	1	-0.118
PR1YR	-0.157	0.105	-0.118	1

Panel B reports the standard deviation in percentage for the factors. The factors generate stable standard deviations in almost every period. Only the last period displays factors with a significantly smaller standard deviation, which implies that the volatility of the factors decreased in the last period. The same results are reported in Panel C, which displays the maximum and minimum factor returns. The results show that SMB generated the highest maximum factor return in the first period and that HML generated the lowest maximum factor return in the final period. Panel C also show that MKT generated the highest negative factor return for all the factors in the first period.

Panel D reports the correlation between factors. The results in the matrix display little correlation between the factors except for the negative correlation of -0,448 between the MKT and the SMB factor. This correlation implies that the SMB factor will decrease with almost half the same return when the Market factor return increases. This result is confirmed through the mean factor returns for the different time periods between MKT and SMB, where they change in different directions.

4.5 Potential biases

In performance evaluation of mutual funds, there could arise several potential biases that should be accounted for. In performance evaluation of mutual funds, there is the possibility of survivorship bias, introduced by Brown, Goetzmann, Ibbotson and Ross (1992), where they analyzed the relationship between volatility and returns. Their analysis was performed onto a sample containing possible survivorship biases, and their results pointed towards the appearance of predictability. In essence, the potential loss of representative results when removing short period lasting funds creates the basis for this bias. If funds with a short lifespan were omitted, these funds may potentially possess a high probability of underperformance before their termination. Therefore, the funds with a short life span will have a significant negative impact on the overall fund performance in the market. The market analyses would report a falsely high return by excluding these observations as the worst funds are omitted from the sample. In addition, these funds are assumed to have a poor or failing strategy due to their alleged underperformance. When assessing management performance in the mutual fund market, the fund's strategy is vital in identifying performance. By omitting a fund's return, the strategy itself will be omitted from the analysis (Elton, Gruber and Blake, 1996). In this effect, the remaining funds in the sample would consist of more or less successful strategies only, creating an upward-biased sample.

Our sample has a minimum requirement of 12 observations for each fund and is thus susceptible to look-ahead bias and even survivorship bias. Look-ahead bias may occur when month-end returns are missing or a requirement of observations is imposed. A fund with a short life span may be excluded from the sample due to the sampling conditions, resulting in the loss of potentially underperforming funds in the sample. These funds may be terminated within a short period due to management costs not covering the funds' expenses or being too high to compete. It could also be due to underperformance on the funds' return. Nevertheless, the omitted funds may affect the sample's overall return, creating a bias when omitted. Our minimum requirement of 12 observations may impose look-ahead bias regarding the funds' life span, but not its size. Funds are rarely directly dissolved, often merged with other funds, thus assuming the money from dissolved funds as transferred directly to another fund (Elton et al., 1996). By this assumption, the minimum requirement of 12 observations should not create any look-ahead bias due to the money still in circulation and its losses and gains incorporated by another fund in the sample. The third potential bias may occur when funds perform a pre-release, not open to the public, onto mutual funds databases for seed money, creating incubation bias (Fama & French, 2010). The fund may create an artificially high increase in market share value during and after this period, rendering the fund's returns disproportionate to the overall sample of funds.

Table 5: Descriptive statistics of fund return and benchmark

The table reports various descriptive statistics for the fund returns and the benchmark for a equally weighted portfolio, consisting of Oslo Børs All Share Index(OSEAX), all funds in the sample, the surviving funds today (alive) and the non-surviving funds today (dead). All numbers are in percentages (%). The two first columns report the mean return and standard deviation annualized, following the kurtosis (not excess) and skewness of each portfolio and finally the maximum and minimum are annualized. In Panel A, we display the whole sample period. Panel B, C and D display time periods at 11 years interval; 1987-1997, 1998-2008 and 2009-2019.

	Mean ret.	Std.dev	Kurt.	Skew.	Max.	Min.
Panel A: 1987-2019						
OSEAX	12.378	20.618	2.836	-0.987	17.445	-27.423
All	12.597	20.635	2.113	-0.812	17.546	-25.277
Alive	13.709	20.808	2.042	-0.793	18.648	-25.391
Dead	10.916	20.967	2.117	-0.793	18.145	-25.088
Panel B: 1987-1997						
OSEAX	16.459	22.620	2.623	-1.061	17.445	-27.423
All	17.848	21.066	0.955	-0.572	17.546	-19.560
Alive	19.813	21.534	0.826	-0.599	18.648	-17.031
Dead	16.225	21.395	1.395	-0.601	18.145	-21.477
Panel C: 1998-2008						
OSEAX	7.489	23.990	1.683	-0.984	12.489	-23.934
All	5.378	24.934	1.427	-0.951	13.821	-25.277
Alive	6.363	24.993	1.443	-0.922	16.112	-25.391
Dead	4.407	25.024	1.422	-0.957	14.102	-25.088
Panel D: 2009-2019						
OSEAX	13.186	13.827	0.914	0.086	15.047	-8.841
All	14.566	14.540	1.597	0.016	15.541	-10.401
Alive	14.951	14.468	1.611	0.030	15.616	-10.324
Dead	12.144	15.192	1.269	0.045	15.417	-10.581

To address the potential biases, we will display fund returns with segregation between alive and dead funds. In Table 5 we have displayed the descriptive statistics for all funds, whether surviving and non-surviving, as well as the benchmark OSEAX. The results are displayed periodically in intervals of 11 years and the overall summary from 1987-2019. For the 11 year periods and all-time, the mean return of non-surviving funds is consistently lower than the mean return of the benchmark index. In contrast, surviving funds are consistently higher, except for

1998-2008. Panel C displays the period of 1998-2008, which contain the worldwide financial crisis. This is reflected in the mean returns of the different portfolios, where the benchmark is outperforming the others. The other panels present the alive funds exceeding the other portfolios and the dead funds underperforming. This is consistent with our previous assumption that non-surviving funds are mainly dissolved due to underperformance.

Figure 4: Cumulative returns

Figure 4 shows the periodically changes in benchmark (OSEAX), all funds, alive funds and dead funds. Panels B, C and D depict the changes in an 11 year window, and the graph represent the overall cumulative return of the different variables. Panel A represent the whole time period from 1987-2019. The method of periodically segregating the development, allow the assessment of the differences in growth at a shorter time span to produce more insight into substantial differences between the factors.

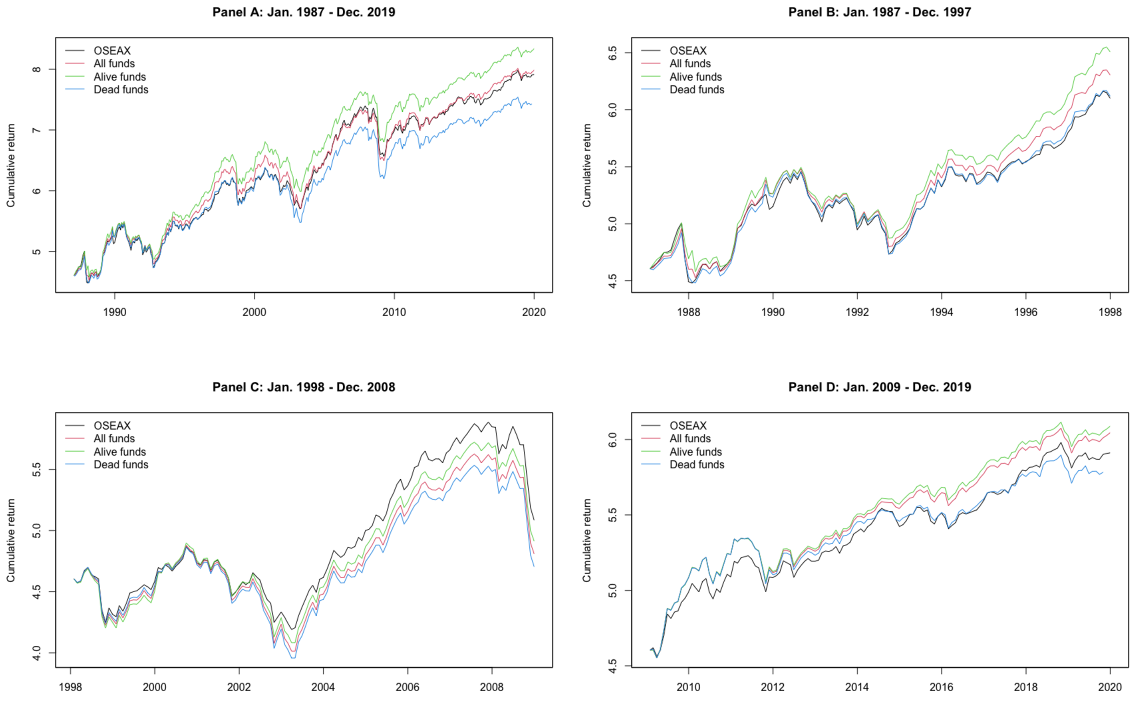


Figure 4 display four sets of graphical development of the cumulative returns in the time periods, illustrating the development of OSEAX, alive funds, dead funds and all funds. Cumulative return is the aggregated change in return over a given period of time. Panel A displays an overall higher development of cumulative returns of alive funds compared to the others. This tendency is also visible in Panels B and D. Panel C (1998-2008) is the exception, as previously stated. Dead funds are persistently lower in cumulative return compared to alive funds and all funds, respectively. With survivorship bias in mind, the omittance of these non-surviving funds would affect our sample upward-biased. Our sample includes both surviving and non-surviving funds, thus combating the survivorship bias.

By our requirement on a minimum of 12 observations, we combat the possibility of survivorship bias and look-ahead bias. Our basis for setting the minimum at 12 is to create a robust sample with a representative selection of funds. If a fund has less than 12 observations, we assume that the fund has gone through troubles of extraordinary measures and is thus not representative of the overall sample. We have also gathered information about limitations from previous work to assess what will be the optimal requirement. Fama & French (2010) have 8 observations, and Sørensen (2011) has a minimum of 17 in his sample. Kosowski et al. (2006) found that observations did not significantly differ for the results when changing from 18 to 120 observations. Fama & French (2010) argue that funds with less than 8 observations have extreme volatility in their respective alphas. Based upon previous work, we operate with 12 observations as our minimum requirement.

5 Empirical analysis

In this section, we present our empirical results on the performance of Norwegian mutual funds. First, we present the results on the aggregate level using the CAPM model (Jensen, 1968), the three-factor model (Fama & French, 1993) and the four-factor model (Carhart, 1997). Second, we use the bootstrap methods of Kosowski et al. (2006) and Fama & French (2010) to determine whether the fund managers' performance is due to skill or luck. Finally, we conduct the double bootstrap method by Harvey & Liu (2020) to estimate the Type II error and assess the Test power.

5.1 Aggregate fund performance

The choice of model depends upon the significance of the factors in the different performance models. By comparing the factors in an OLS regression analysis, we receive the aggregated level of the equally weighted portfolio and the following statistical significance level for each factor. We have divided the analysis into time periods to see which performance model receives the highest statistical significance for each time period. The regression analysis displays the funds' return dependant upon the factors, differing from one to four factors, depending upon the method used. We will focus on the three models featured in the methods of Kosowski et al. (2006), Fama & French (1993) and Harvey & Liu (2020). Table 6 display the models CAPM, Fama & French (2010) three-factor model and Carhart (1997) four-factor model with their respective selection of explanatory parameters and the following measure of significance provided in the parentheses. The models differ in factors where we include the Market factor (MKT), Small-Minus-Big

(SMB), High-Minus-Low (HML) and the Momentum factor (PR1YR) varying between the models.

In Table 6 we see the results of the regression. In Panel A, CAPM displays an alpha of 0.501 with a t-statistic of 0.485, which is nonsignificant. However, the Market factor produces a high significance with a t-statistic of 67.10. The alphas in Panel A for all three models do not generate statistical significance in their t-statistics, thus revealing that none of the listed models can provide an alpha with high statistical significance of abnormally high returns over 1987-2019 on the aggregated level. Further, the Fama & French (1993) three-factor model includes the SMB and HML in addition to the alpha and the Market factor (MKT). Panel A displays a decrease in alpha to -0.379 for the CAPM with a nonsignificant t-statistic. However, the factors of this model display statistical significance at the 1 percent level. The Carhart (1997) four-factor model further includes Momentum (PR1YR) and displays an alpha of 0.035. MKT, SMB and HML factors produce significance at the 1 percent level, while the Momentum factor presents significance at the 5 percent level. The increase in alpha from Fama & French to Carhart is evidence of positive growth in alpha by adding the Momentum factor, increasing funds' overall aggregated return over the time period.

In the different time periods, the models' values vary greatly in impact and statistical significance. Panel B displays significance at 1 percent for the MKT and SMB in all models. The HML and PR1YR does not generate statistical significance at any level below 10 percent and produce low levels of statistical significance for the funds' aggregated return in 1987-1997. In addition, the alphas remain statistically nonsignificant. Panel C returns overall negative alphas for all methods. The negative alpha from Fama & French's (1993) three-factor model is the only statistically significant alpha in the table. It depicts an extensive underperformance in the Norwegian mutual funds market for this method in the time period 1998-2008. All factors display significance at the 1 percent level during this time period and are seen as significantly explanatory towards the negative alphas received. The time period ends in 2008 and is thus affected by the financial crisis at the end of the time period. Panel D contains overall positive alphas, with CAPM producing the highest. None of these models generate statistically significant alphas. However, we find that the MKT factor is robust at a 1 percent level for all methods.

In summary, we see that the CAPM model produces high alpha values either negatively or positively in all panels. By only including one explanatory variable, the alpha seems to generate higher values. By adding more factors, we see the alpha moving further towards zero in all time periods. In this effect, the more factors we include, the less significance and impact from the alpha on the funds aggregated returns. In the choice of benchmark method, significant impact values are of high importance to substantially explain the development in the market. For Panel

A 1987-2019, the Carhart four-factor model display significance for all four factors for the whole time period. In this effect, the Carhart model is the preferred approach to provide the most robust regressional analysis with the highest significance of impact towards the alpha.

Table 6: Equally weighted portfolio alphas and factor loadings

The table displays the aggreagate performance where the columns present the alpha, factor loadings and adjusted R square for an equally weighted portfolio. We present the results for the three different factor models; CAPM, Fama & French and Carhart. The stars represent the statistical significance 1%***,5%** and 10%* and the alphas are annualized and in percent. For Panel A display results for the whole time period. Panel B-D displays equally sub periods; 1987-1997, 1998-2008 and 2009-2019.

	α	βMKT	βSMB	βHML	$\beta PR1YR$	${R^2}_{adj}$
Panel A:Jan 1987-Dec2019						
CAPM	0.501	0.960***				0.919
	(0.485)	(67.10)				
Fama-French	-0.379	0.993***	0.099***	-0.060***		0.926
	(0.372)	(64.538)	(4.338)	(-3.393)		
Carhart	0.035	0.988***	0.099***	-0.065***	-0.044**	0.926
	(0.034)	(64.133)	(4.429)	(-3.662)	(-2.480)	
Panel B: Jan 1987-Dec1997						
CAPM	2.246	0.882***				0.892
	(1.061)	(32.840)				
Fama-French	0.231	0.914***	0.141***	0.032		0.899
	(0.108)	(30.935)	(3.384)	(0.906)		
Carhart	0.146	0.914***	0.142***	0.038	0.022	0.898
	(0.068)	(30.865)	(3.378)	(1.031)	(0.632)	
Panel C: Jan 1998-Dec2008						
CAPM	-2.141	1.012***				0.949
	(-1.247)	(49.190)				
Fama-French	-3.495	1.026***	0.122***	-0.106***		0.958
	(-2.169)	(47.863)	(3.453)	(-4.080)		
Carhart	-2.454	1.018***	0.134***	-0.109***	-0.104***	0.963
	(-1.603)	(50.431)	(4.037)	(-4.444)	(-4.294)	
Panel D: Jan 2009-Dec2019						
CAPM	1.313	1.006***				0.914
	(0.993)	(37.353)				
Fama-French	0.810	1.024***	0.033	-0.053*		0.915
	(0.602)	(32.153)	(0.899)	(-1.635)		
Carhart	0.776	1.025***	0.033	-0.053	0.002	0.914
	(0.537)	(30.319)	(0.898)	(-1.628)	(0.065)	

5.2 Methodology of Kosowski et al. (2006)

The previous section found that the Norwegian mutual fund managers could not produce a statistically significant alpha on an aggregate level. This section will evaluate if any of the fund managers can beat their benchmark index individually. When conducting a multiple test, some funds will perform well, and some will perform poorly in any given sample. The bootstrap procedure in this section will consider lucky or unlucky performances in the sample. The method by Kosowski et al. (2006) allows us to examine individually whether the actual alpha is different from zero in the bootstrapped distribution of alphas by the null hypothesis of zero alpha for all funds. The procedure allows us to distinguish luck from skill and whether skilled or unskilled mutual fund managers exist in the market. We will also compare the results from the bootstrapped p-value with the standard parametric p-value on how the conclusions on significance will differentiate between the two tests.

We will look for statistical significance in the two different performance measures, alpha and the t-statistic of alpha, in all the mutual funds individually. Table 7 represents the results from the bootstrap procedure where we use the alpha α in the first column and t-statistic of alpha t_α in the second column. The funds are ranked by their respective alpha t-statistic, where Panel A; presents the 15 best funds and Panel B; presents the 15 worst funds. The third column contains the parametric p-value of each fund, and the fourth column represents their bootstrapped p-value.

Further, Appendix E presents the 107 Norwegian mutual funds individually with their two separate performance measurements, alpha in row 1 and t-statistic of alpha in row 2. The table also reports the parametric p-value in row 3 and the bootstrapped p-value in row 4. The funds are sorted based on their alpha t-statistic. To evaluate whether the performance of any of the funds is based on luck or stock-picking skills, we look for statistical significance using the alpha and t-statistic of the alpha of the mutual funds. E.g., if we want to evaluate whether the best performing fund displays stock-picking skill in contrast to a lucky performance on the 5% significance level, at least 5% of the generated alphas in the 1000 simulations must be higher than the observed alpha in the four-factor model. Equivalently, for the worst-performing funds, if the performance is due to a lack of stock-picking skills. At least 5% of the generated alphas need to be lower than the observed alpha in the four-factor model rendering the bad performance due to a lack of stock-picking skills.

In Table 7 we can see that the top 15 funds with positive alphas beat their benchmark. If we look closer at the top-performing funds, FIRST Norge Fokus (FNK) display an alpha of 10.95, the 2nd highest fund Landkreditt Utbytte 1 (LU1), display an alpha of 7.62 and the 3rd FORTE Trønder (FTB) with an alpha of 6.18. The top 3 performing funds exhibit nonsignificant alphas as their respective bootstrapped p-values are above the threshold of 0.05. However, LU1 displays a bootstrap p-value of 0.10, lower than the bootstrapped p-value of the other top-performing funds with 0.79 and 0.62, respectively, yet it remains nonsignificant. Furthermore, all the top 15 funds exhibit positive alphas, and we fail to reject the null hypothesis of zero performance for all of them.

When we look at the funds with negative alphas, we can reject the null hypothesis of no true performance on a 5% significance level for all of them, except for the worst-performing fund Skandia SMB Norge (SSM), with an alpha of -12.25. SSM displays a bootstrapped p-value of 0.16, above the threshold of 0.05, indicating no evidence towards a lack of skill. Our result shows that we fail to reject the null hypothesis of zero performance for the managers that generate positive alpha. Their performance resulted from lucky performances rather than the display of stock-picking skills. However, we reject the null hypothesis of zero performance for all managers that generate a negative alpha, except for SSM. The result implies that their bad performances are not a result of unlucky performances rather the absence of skill.

The evaluation of t-statistics of alpha presented in column two provides a conclusion similar to the deduction of the alpha values. The top-performing funds, according to t-statistics of alpha does not display significance in their bootstrapped p-value. Thus their performances remain, as previously concluded, a result of luck. The 7th best performing fund (LU1) has the lowest bootstrapped p-value of 0.10. This value is still nonsignificant on a 5% level, failing to reject the null hypothesis. Our results still conclude that the performance of all the funds exhibiting a positive alpha t-statistic fails to reject the null hypothesis.

Among the funds that fail to beat their benchmark and exhibit negative t-statistics of alpha, the conclusion is the same as for alphas. All funds with negative alpha t-statistics, except for one fund, reject the null hypothesis. The bottom fund (SSM), which now is the worst-performing fund with a t-statistic of alpha at -3.22, generates a bootstrapped p-value of 0.16, failing to reject the null. This result indicates that their performance is rather unlucky, not by a lack of stock-picking skills. Thus, based on the evidence, all positive t-statistics of alpha fail to reject the null hypothesis indicating no evidence of stock-picking skills. In contrast, we can reject the null hypothesis for the mutual funds that generate a negative t-statistic of alpha, implying that the bad performances are due to a lack of stock-picking skills.

Table 7: Baseline bootstrapped results

The table below present the individual bootstrapped results from Kosowski et al. (2006) baseline bootstrap. The results presented in Panel A for the top 15 funds and Panel B for the bottom 15 funds ranked on their alpha t-statistic. Column two present the estimated alpha for each fund, column three; the ranked alpha t-statistics, column four; the parametric p-value and column five; the bootstrapped p-value.

	α	α_t	p-value	Boot p-value
Panel A: Top 15 funds				
Danske Invest Norge Aksj. Inst 1	3.02	2.23	0.01	0.80
FIRST Norge Fokus	10.95	2.19	0.01	0.50
Landkreditt Utbytte	5.38	2.15	0.02	0.35
Fondsfinans Norge	3.78	1.94	0.03	0.37
Danske Invest Norge Aksj. Inst 2	3.26	1.94	0.03	0.21
PLUSS Markedsverdi (Fondsforv)	1.95	1.83	0.03	0.21
Landkreditt Utbytte I	7.62	1.82	0.03	0.10
FORTE Trønder	6.18	1.43	0.08	0.66
Storebrand Norge I	1.60	1.36	0.09	0.64
Storebrand Norge Fossilfri	4.18	1.24	0.11	0.78
Storebrand Optima Norge	1.72	1.19	0.12	0.77
Landkreditt Norge	3.19	1.15	0.12	0.69
Carnegie Aksje Norge	1.78	1.15	0.13	0.62
DNB Norge R	6.11	1.11	0.13	0.55
Danske Invest Norge II	1.39	1.07	0.14	0.60
Panel B: Bottom 15 funds				
DNB Norge	-1.21	-1.44	0.08	0.03
DNB Norge (Avanse II)	-1.72	-1.53	0.08	0.03
Fokus Barnespar	-11.89	-1.67	0.07	0.01
Alfred Berg Vekst	-8.95	-1.72	0.06	0.01
Globus Norge II	-8.34	-1.73	0.05	0.02
GAMBAK Oppkjøp	-18.62	-1.81	0.04	0.02
Danske Invest Aktiv Formuesf. A	-19.06	-1.89	0.04	0.02
Globus Norge	-8.54	-1.99	0.04	0.01
Alfred Berg Aksjef Norge	-3.32	-2.03	0.03	0.02
Alfred Berg Aksjespar	-4.98	-2.14	0.02	0.02
GJENSIDIGE AksjeSpar	-4.28	-2.45	0.01	0.00
Nordea SMB	-6.44	-2.63	0.00	0.01
GJENSIDIGE Invest	-5.31	-2.96	0.00	0.00
Nordea SMB II	-16.93	-3.18	0.00	0.02
Skandia SMB Norge	-12.25	-3.22	0.00	0.16

The result would be different if we used the generated parametric p-value instead of the bootstrap p-value to test the null hypotheses. Some of the top-performing mutual funds ranked on their alpha t-statistic display parametric p-values below 5%. We would reject the null hypothesis, implying that these mutual fund managers possess stock-picking skills. In Table 7, seven of the funds exhibiting positive alpha t-statistics have display p-values below 0.05. Implying that we would reject the null hypothesis of zero performance for outperforming managers, resulting in Norwegian mutual fund managers with stock-picking skills through the traditional method of the parametric p-value.

When we use the parametric p-value to assess the underperforming funds ranked on their t-statistics of alpha, the results deviate from the bootstrap test. Some of the worst-performing funds exhibit parametric p-values above 0.05, implying zero performance for the worst-performing funds. These results would imply that the bad performance of mutual fund managers is not a consequence of bad stock-picking skills, rather unlucky performances. We would thus experience a contradiction towards the results from the bootstrap baseline, which illustrates the importance of the procedure developed by Kosowski et al. (2006).

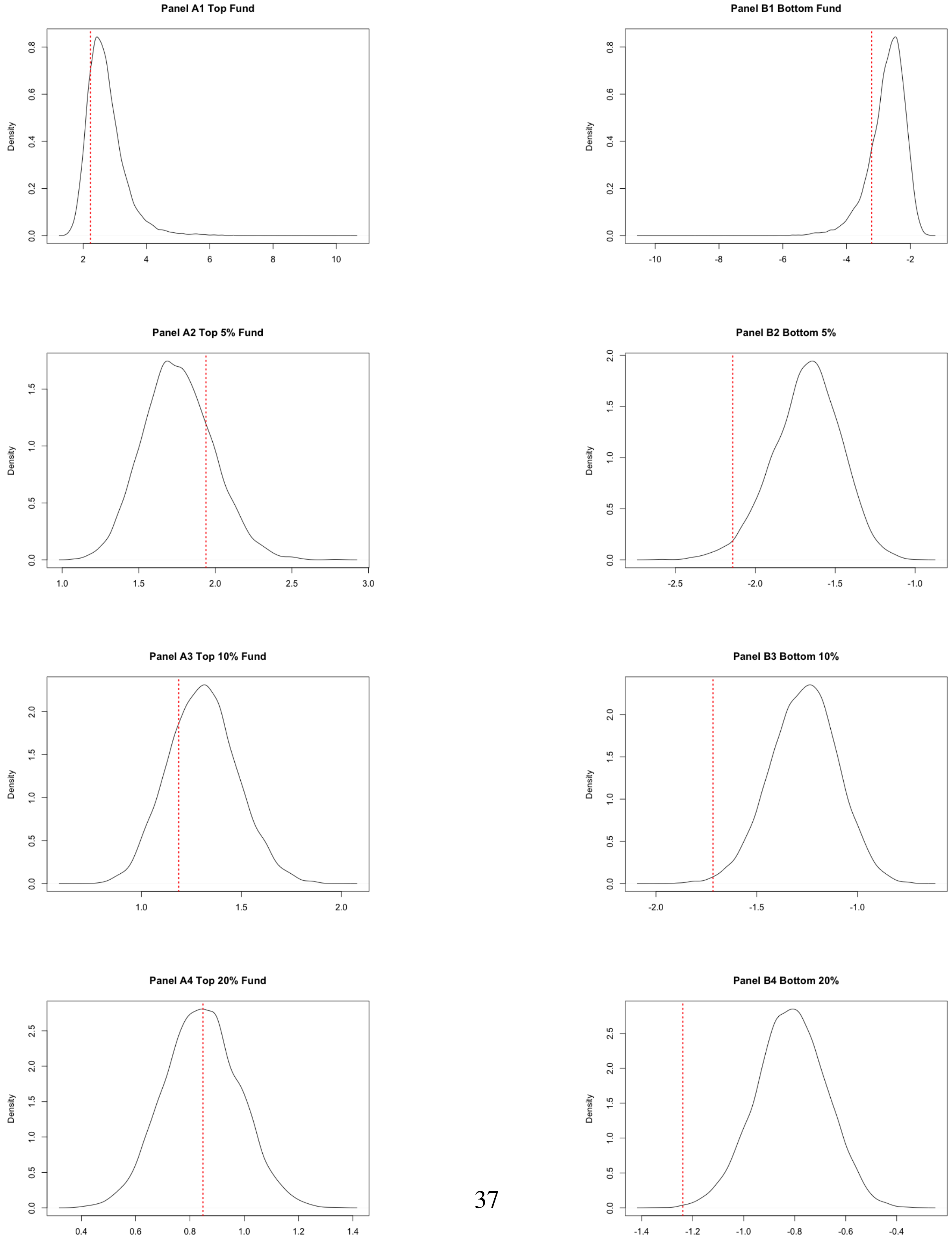
Figure 5 present the distribution of unconditional four-factor alphas for the top performing 5%, 10% and 20% of funds on the left side of the panel. The bottom performing 5%, 10% and 20% are displayed on the right side. The dashed red line represents the previously estimated t-statistic of alpha. Panel A1 displays the distribution of the top-performing fund (Danske Invest Norge Aksj. Inst 1). From the distribution presented in the graph, the probability of observing a t-statistic of 2.23 is very likely. Which implies that we would not be able to reject the null hypothesis for this fund. Thus, the performance is based on luck rather than skills.

The distribution displayed in Figure 5 present the alpha t-statistic of a specific mutual fund, generated under the null hypothesis of zero performance in our benchmark factor model and simulated 1000 times. We can only reject the null hypothesis if the original alpha t-statistic is unlikely (less than 5% chance) to be simulated. As a result, we can see that the rest of the alpha t-statistics in Panel A are likely to be observed in the distributions. Therefore, we are not able to reject the null hypothesis for these panels.

Panel B1 represents the bottom fund with an estimated alpha t-statistic of -3.22. The estimated alpha t-statistics lie within the simulated distribution, and we fail to reject the null hypothesis. As previously demonstrated, the worst performing fund is not a case of bad stock-picking skills but an unlucky performance. Panel B2 displays the distribution of simulated alpha t-statistics that lie approximately between -1.8 and -1. The observed alpha t-statistic of -1.73 is within the 5% rejection area, and we can reject the null hypothesis. This result is the same for Panel B3 and B4 and is representative for the rest of the funds with negative alpha t-statistics.

Figure 5: Bootstrapped distribution vs estimated t-statistic of alpha

The panels display the Kernel density of alpha distribution when using bootstrapped unconditional four-factor t-statistics represented by the solid line. The dashed red line represent the estimated funds t-statistic. Panel A reports the right tail of the distribution as Panel B reports the left tail of the distribution. The statistics are computed from 1000 bootstrap resamples and ranked on their t-statistics of alpha.



Panel A1 Top Fund

Panel B1 Bottom Fund

Panel A2 Top 5% Fund

Panel B2 Bottom 5%

Panel A3 Top 10% Fund

Panel B3 Bottom 10%

Panel A4 Top 20% Fund

Panel B4 Bottom 20%

Panel B1-B4 all reject the null hypothesis with bootstrapped p-values lower than 0.05. Given that only one of the 62 funds that exhibit a negative alpha t-statistic fails to reject the null hypothesis, the evidence towards lack of skill among the underperforming funds is clear. The evidence agree with the results of both Kosowski et al. (2006) and Sørensen (2011); an absence of skill causes abnormal negative performance.

5.3 Methodology of Fama & French (2010)

We evaluated if any of the 107 mutual fund managers individually display skills to beat their benchmark model in the previous method. We got the same result as in the aggregate performance. We cannot find any statistically significant evidence that any mutual fund managers beat their benchmark through stock-picking skills. The methodology by Fama & French (2010) allows us to evaluate statistical significance for the best and worst-performing funds ranked on their t-statistic in a joint test. We follow the same conditions as the previous test, allowing us to test for significance under zero alpha for the specific funds. We use the Carhart (1997) four-factor model as our factor benchmark model. The test evaluates the 5 best and 5 worst funds ranked by alpha t-statistics. We have jointly sampled the group of funds and their explanatory returns. We describe the methodology of the procedure in Section 3. The procedure lets us evaluate whether the grouped best or worst mutual fund managers exhibit skill or luck in their performances. At the same time, we consider the possibility of correlation between the simulated alphas. Such correlation may be a result of the benchmark factor model failing to capture all the external variations.

Table 8: Bootstrap jointly result

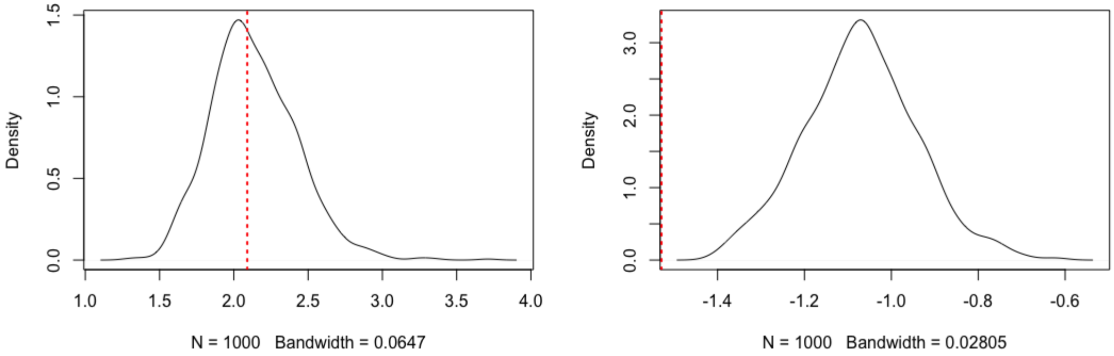
The table present the average alpha t-statistic for the 5 best and 5 worst funds ranked on their t-static. The bootstrapped p-value is jointly sampled for all the 5 funds to account for correlation in the alpha.

	Best funds	Worst funds
Average $\alpha t - statistic$	2.09	-1.53
Bootstrapped p-value	0.484	0.004

In this test, we evaluate the t-statistics of alphas. Fama & French (2010) promote t-statistics rather than the estimated alpha, as it considers the precision with which alpha is estimated. Table 8 tells us that the bootstrapped p-value for the 5 best funds is not significant with a p-value equal to 0.484, far above the required significance level of 0.05 for p-value significance. As a result, we are not able to reject the null hypothesis of zero alpha. On the other hand, the average t-statistic of 2.09 is pretty likely to be observed when the null hypothesis is true, as we can see from Figure 6 where the estimated t-statistic is located close to the center of the distribution. The result further indicates that the alpha t-statistic originates from luck rather than skill. We are still not able to detect skills among managers.

Figure 6: Bootstrapped jointly distribution

The figure shows the density distribution of average alpha t-statistics for the top 5 best performing funds and the bottom 5 worst performing funds in the first bootstrap. The distribution displays the average alpha t-stat for the 5 funds in question. The figure on the left contains the alpha t-stat observations of the top 5 best funds, and the red line represent the alpha of the top 1 best performing fund from the ranked t-statistics table. The figure on the right contains the distribution of the top 5 worst performing funds alpha t-stat. The number of observations is 1000 for both graphs due to the bootstrap of 1000 simulations in the analysis. The red line representing the worst 1 performing fund is displayed far to the left of the figure and appears outside the distribution.



Further, Table 8 shows the bootstrapped p-value for the 5 worst funds equal to 0.004, which is highly significant. Hence this alpha t-statistic is unlikely to be observed if the null hypothesis is true and we can reject the null hypothesis. We could easily reject the null hypothesis based on the density distribution in Figure 6 by the average alpha t-statistic of -1.53. The result contributes with evidence, similarly to the previous tests, that the worst mutual fund managers lack skill, and the poor result is not a result of unlucky performances. A problem with this joint test is that the worst-performing fund based on alpha t-statistic, Skandia SMB Norge, is wrongly convicted of being unskilled. In Figure 7 they fail to exhibit a significant bootstrapped p-value, indicating that their performance was unlucky. However, the results contribute to our previous test as we can detect underperforming managers. The results coincide with Sørensen's (2009) on Norwegian mutual funds and Fama & French's (2010) on US mutual funds.

Figure 7: Simulated vs actual alpha t-statistic

The figure illustrate the simulated alpha t-statistics from our bootstrap approach vs the actual alpha t-statistic. The two vertical lines illustrates the bottom and top 5 funds in the distribution. The four-factor model is used to compute the t-statistics of alphas for the entire sample period; 1987-2019

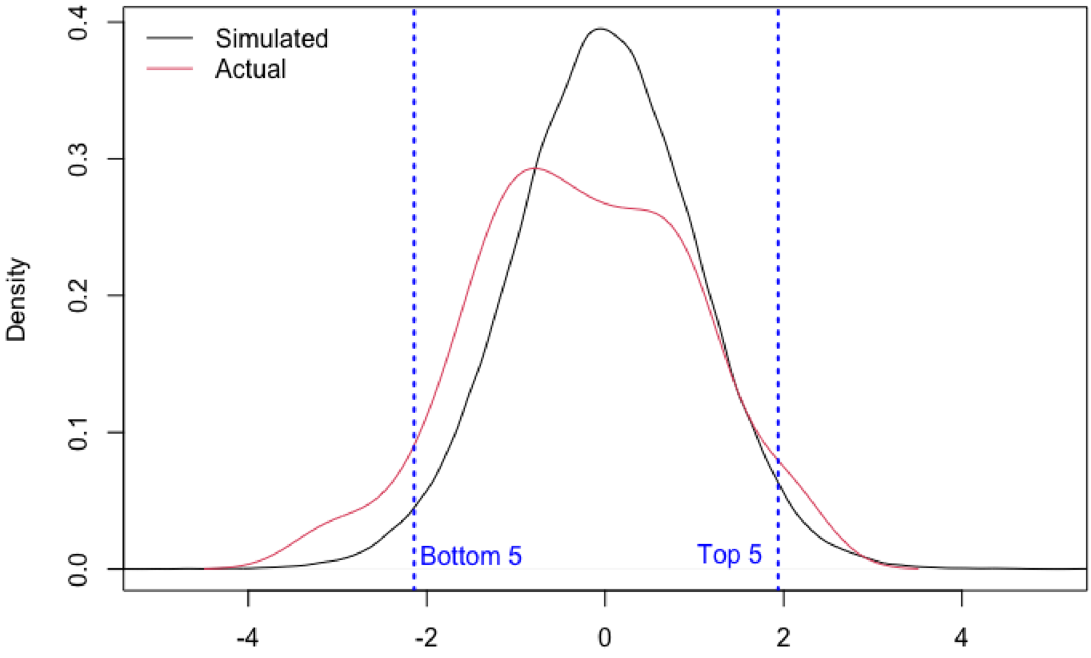


Figure 7 compares the density of the simulated alpha t-statistics from our bootstrap approach versus the generated alpha t-statistic. The actual t-statistic is estimated in the factor model for each fund without the assumption of zero performance, from Equation 4. The simulated distribution display more mass of probability on the t-statistics close to zero, whereas the generated t-statistic of alpha display more mass of probability in the tails. The vertical lines distinguish the bottom and top 5 funds, ranked on their alpha t-statistic. For these funds, the simulated distribution has thinner tails than the actual alpha t-statistics. This is because the actual performance features ”shoulders” in these regions under the normality assumption. According to Kosowski et al. (2006), the bootstrap inference does not only measure fat or thin tails of the actual distribution but also captures the complex shape of the entire distribution of t-statistics under the null hypothesis.

As we can see from Figure 7, the distribution display somewhat thinner tails on the right side of the distribution. As a result, the bootstrap approach as presented has a better explanatory effect on the positive alpha t-statistics than the negative alpha t-statistics.

5.4 Methodology of Harvey & Liu (2020)

Until now, we have only considered the possibility of selecting a mutual fund manager that turns out to be unskilled (Type I error). However, we have not considered the possibility of missing a manager that turns out to be skilled (Type II error). The methodology of Harvey & Liu (2020) consider this error rate that allows us to estimate the Test power. First, we will implement their bootstrap methodology to estimate the cutoff t-statistic for various alternative hypotheses to control for luck. Second, we use the double bootstrap approach that allows us to estimate the Type II error rate, Oratio and calculate the Test power.

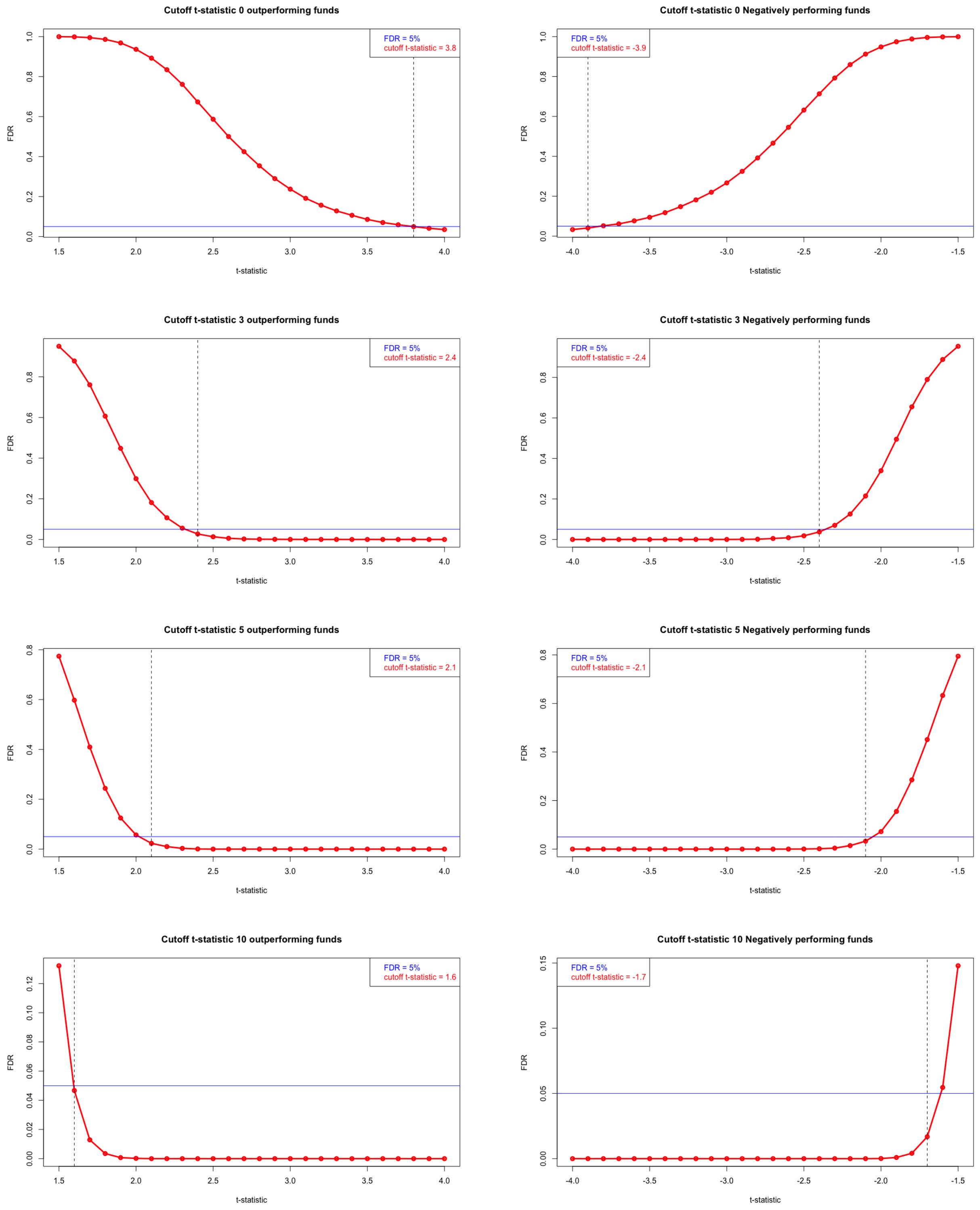
5.4.1 Cutoff t-statistics

In the first part, we estimate the cutoff t-statistic corresponding to a false discovery rate at 5%. We operate under the null hypothesis of no true performance $\alpha = 0$ for all funds and estimate the cutoff for various alternative hypotheses H_A^N as presented in Table 1. The Type I error rate is associated with the FDR and will be the error rate we estimate in this test. Harvey & Liu (2020) argue that the cutoff t-statistic method is essential when assessing significance in the dataset. They argue that a pre-defined cutoff t-statistic, commonly 2.0 or 3.0 from Harvey, Liu and Zhu (2016), will be misguiding for the actual threshold of the sample. Using a hurdle of t-statistics from 1.5 to 4.0, we perform a reversed approach. The funds' alpha t-statistic distribution determines the cutoff of significance when controlling for a specific level of FDR, making the estimated cutoff data-specific. The method of cutoff t-statistics for zero outperforming or underperforming funds is aligned with Kosowski et al. (2006) and Fama & French (1993). The method of Harvey & Liu (2020), as described above, will provide aligned results as we will demonstrate.

We will examine both tails of the distribution to assess whether there exists skill among managers, under the assumption of zero performance. In Appendix F, we estimate the FDR for each alternative hypothesis H_A^N, in a cutoff sequence from 1.5 to 4.0 (or -4.0 to -1.5) by 0.1 increments. The cutoff t-statistic is defined for each hypothesis as the first observed cutoff increment where $FDR < 0.05$. The cutoff t-statistics are closely similar for both positive and negative cutoff t-statistics, as shown in Appendix F.

Figure 8: Optimal cutoff t-statistic distribution

Figure 8 shows the simulated t-statistic of alphas for a selection of H_A. It depicts the cutoff t-statistic where we control for a 5 percent False Discovery Rate (FDR). The graphs on the left show distribution of outperforming funds, and underperforming funds to the right.

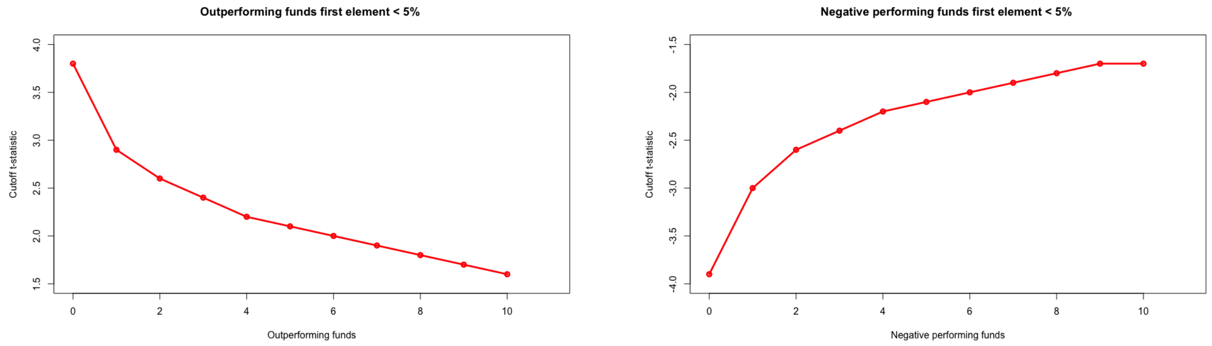


The first method serves to find the cutoff t-statistic for a given H_A in the sample under the null hypothesis of zero performance for all funds. From Figure 8 the distribution of t-statistics for the respective alphas dependant upon the FDR is displayed for a selection of H_A. As an example from Figure 8 a), where $H_A^3 > 0$, we test whether the three best funds display positive alphas. Under this hypothesis, the optimal cutoff when controlling for a 5% false discovery rate is found at 2.4. Thus, the cutoff t-statistic represents the level of t-statistic required to reject the null hypothesis (H_0).

Figure 9 presents the cutoff t-statistic as a function of H_A, the number of funds we test to be out-or underperforming, out of all mutual funds. The results in Figure 9 show that the cutoff t-statistic decreases as more funds are believed to be outperforming. This result is, according to Harvey & Liu (2020), because discovery is less likely to be false when a larger fraction of the data is true. This corresponds to Figure 9 which shows that when we test for 1 fund outperforming, the cutoff t-statistic is higher (2.9) than if a test for 2 funds outperforming (2.6).

Figure 9: Cutoff t-statistic as function of H$_A$

The figure display the out- and underperforming cutoffs at first observed element of $FDR < 0.05$. It illustrates the cutoffs differing on the different H_A. The graph on the left display the development of cutoffs for outperforming funds, and on the right, underperforming funds.



Through this method, we can test the null hypothesis of zero performance for all funds under the assumption of different alternative hypotheses H_A. Using the bootstrapped sample of the funds' return, we obtain a t-statistic threshold for different False Discovery Rate values (FDR). The cutoff t-statistic for each hypothesis is compared to the generated alpha t-statistic for each fund in the method of Kosowski et al. (2006). If any funds display a higher t-statistic of alpha than the cutoff, H_0 must be rejected as the assumption of zero funds outperforming cannot hold true. As presented in the methodology, we will operate under the hypotheses as shown in Table 1 where we test for $N = 10$ funds assumed to have alpha values larger ($\alpha > 0$) or smaller ($\alpha < 0$) than zero.

Table 9 provides a summary of the overall results of our cutoff t-statistic analysis using the dataset of all funds with $\alpha = 0$. The table is divided into top and bottom performing funds; Panel A presents the results from the analysis with the Top 10 performing funds, ranked by alpha t-statistic from the procedure of Kosowski et al. (2006). Panel B displays the same results for the 10 bottom-performing funds. The table display statistical results regarding the individual fund's t-statistic of alpha, the cutoff t-statistic for a given alternative hypothesis and whether the null hypothesis can be rejected for the H_A in question. From Table 9 for H_A^0, we can see the cutoff t-statistic, at 3.8, being much higher than any fund's alpha t-statistic from the procedure of Kosowski et al. (2006), which implies that the assumption of zero outperforming funds may, or may not be true, as there is no suitable significance in the sample. Thus, we cannot reject the H_0 of zero funds outperforming. Next, we will look at some examples from the table.

For H_A^1, the estimated cutoff t-statistic is 2.9. For the H_0 to be rejected, there needs to be one fund displaying a t-statistic of alpha higher than the cutoff. The best performing fund's t-statistic of alpha is 2.23, and we see that no fund beats the threshold, and thus the H_0 cannot be rejected. For H_A^5, the cutoff is 2.1, and to reject the H_0, at least 5 funds must display a higher alpha t-statistic than the cutoff. The fifth best fund displays a 1.94 alpha t-statistic, and since the table is ranked, we can still not reject the H_0 of no funds outperforming. These two examples are representative of all the top funds in our analysis. For all p_0, the H_0 can not be rejected, and we fail to find evidence towards skillful managers in the sample with statistical significance.

The tendency towards more funds being statistically significant in underperformance due to a lack of stock-picking skill is apparent. We see that H_A^0 for underperforming funds does not provide any funds with a lower alpha t-statistic, so we cannot reject H_0. However, for H_A^1, the cutoff t-statistic is at -2.6. Therefore, to reject the H_0, we must identify at least one fund with a lower alpha t-statistic. We see that the worst-performing fund displays an alpha t-statistic of -3.22, being above the threshold. Thus, we can reject H_0 of no funds underperforming and accept the H_A^1 of at least one fund underperforming. For H_A^7, the cutoff is -1.8, and we must identify at least 7 funds with an alpha t-statistic below this cutoff. The seventh worst performing fund displays an alpha t-statistic of -2.03, and we can reject H_0 of no funds underperforming, accepting the H_A^7 of at least seven funds underperforming. All alternative hypotheses for underperforming funds are accepted as we consistently identify that funds display alpha t-statistic below the cutoff.

For Table 9, we return to finding a lack of statistical significance in locating outperforming funds. We can not reject H_0 for any H_A and can not, with any statistical certainty, find evidence towards any funds outperforming due to skillful managers. For the underperforming funds, we identify significance alpha t-statistics, being able to reject H_0 for every H_A tested for. The results show us that we can not identify any outperforming funds explained by skill. Furthermore, that bad performing funds are underperforming due to a lack of skill by using the cutoff t-statistic presented by Harvey & Liu (2020).

Table 9: Cutoff t-statistic test results

The table displays the summary of cutoff t-statistics on a range of H_A. Through the cutoff t-statistic, we measure each hypothesis towards the ranked t-statistic of each fund's alpha and thus determine whether the H_0 is rejected or not. The table also displays the number of observations with a t-statistic higher than the cutoff t-statistic of a particular p_0.

		α_t	Cutoff t-stat	H_A	H_0 reject?
H_A	**Panel A: Top 10 funds**				
0	Zero funds outperform		3.8	$\alpha > 0$	No
1	Danske Invest Norge Aksj. Inst 1	2.23	2.9	$\alpha_1 > 0$	No
2	FIRST Norge Fokus	2.19	2.6	$\alpha_{1,2} > 0$	No
3	Landkreditt Utbytte	2.15	2.4	$\alpha_{1,2,3} > 0$	No
4	Fondsfinans Norge	1.94	2.2	$\alpha_{1,2,..,4} > 0$	No
5	Danske Invest Norge Aksj. Inst 2	1.94	2.1	$\alpha_{1,2,..,5} > 0$	No
6	PLUSS Markedsverdi (Fondsforv)	1.83	2.0	$\alpha_{1,2,..,6} > 0$	No
7	Landkreditt Utbytte I	1.82	1.9	$\alpha_{1,2,..,7} > 0$	No
8	FORTE Trønder	1.43	1.8	$\alpha_{1,2,..,8} > 0$	No
9	Storebrand Norge I	1.36	1.7	$\alpha_{1,2,..,9} > 0$	No
10	Storebrand Norge Fossilfri	1.24	1.6	$\alpha_{1,2,..,10} > 0$	No
	Panel B: Bottom 10 funds				
10	GAMBAK Oppkjøp	-1.81	-1.6	$\alpha_{1,2,..,10} < 0$	**Yes**
9	Danske Invest Aktiv Formuesf. A	-1.89	-1.7	$\alpha_{1,2,..,9} < 0$	**Yes**
8	Globus Norge	-1.99	-1.7	$\alpha_{1,2,..,8} < 0$	**Yes**
7	Alfred Berg Aksjef Norge	-2.03	-1.8	$\alpha_{1,2,..,7} < 0$	**Yes**
6	Alfred Berg Aksjespar	-2.14	-1.9	$\alpha_{1,2,..,6} < 0$	**Yes**
5	GJENSIDIGE AksjeSpar	-2.45	-2.0	$\alpha_{1,2,..,5} < 0$	**Yes**
4	Nordea SMB	-2.63	-2.1	$\alpha_{1,2,..,4} < 0$	**Yes**
3	GJENSIDIGE Invest	-2.96	-2.2	$\alpha_{1,2,3} < 0$	**Yes**
2	Nordea SMB II	-3.18	-2.4	$\alpha_{1,2} < 0$	**Yes**
1	Skandia SMB Norge	-3.22	-2.6	$\alpha_1 < 0$	**Yes**
0	Zero fund under-perform		-3.9	$\alpha < 0$	No

5.4.2 Type II error rate and Test power

In the previous tests, we worked under the assumption of zero performance for all the managers and tested this null hypothesis. However, these methods only allow us to evaluate the Type I error rate. This section extends our analysis, assuming that a fraction of managers outperforms or under-perform, that $H_0 : \alpha \neq 0$. Harvey & Liu (2020) present a double bootstrap procedure under this null hypothesis, allowing us to estimate the Type II error rate. As mentioned in the methodology section, we use the standard approach to estimate the Type II error to calculate the Test power as well.

Figure 10 displays the distribution of alpha t-statistics of our sample and the receiving operating curve (ROC). We will evaluate the Test power of both tails. We can see that the distribution of alpha t-statistics is skewed to the left, contributing to our previous findings that there is more evidence towards lack of skill among underperforming managers than skill among outperforming managers. There are some managers with alpha t-statistics between 1.5 and 4.0 in the distribution, which we want to investigate further. The ROC represents the relationship between the true positive rate and the false positive rate. We can estimate these error rates for a given p_0 in the double bootstrap procedure and simulate it many more times to find the average across all the simulations. The formula for each cutoff, k, is

$$TPR_k = mean\frac{TP_j}{p_0} \tag{21}$$

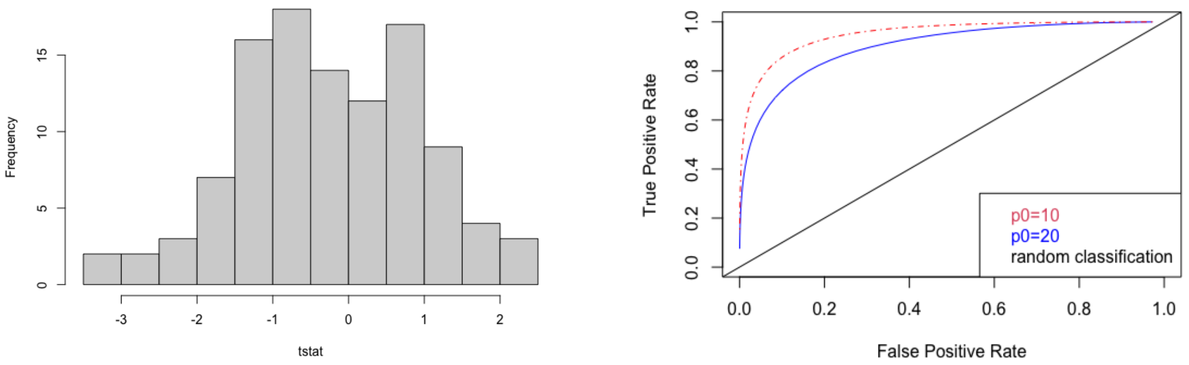
and

$$FPR_k = mean\frac{FP_j}{n - p_0}, \tag{22}$$

to estimate the true positive rate and false-positive rate. Where n represents all the mutual funds and j is the bootstrap iterations. The ideal classification outcome for TPR and FPR is given by the point (0.1) where we have a false positive rate equal to zero and a true positive rate of 100%. This result indicates that a curve closer to this point is deemed better than a curve close to the random classification line. Figure 10 b) displays that the assumption on $p_0 = 10$ is better than $p_0 = 20$. This result can be explained by the fact that a smaller p_0 results in a higher average t-statistic.

Figure 10: Mutual fund performance

These figures display the graphical representation of the Norwegian mutual fund's performance with different descriptive statistics. Figure a) display the distribution of all funds' alpha t-statistics from the Kosowski et al. (2006) method computed from the Carhart (1997) four-factor model and show all 107 funds' t-statistic of alpha. Figure b) is the Receiver Operating Characteristics for all funds from the Harvey & Liu (2020) double bootstrap method for error rate computation. It depicts the relationship between the number of true positives and false positives for $p_0 = 10$ and $p_0 = 20$, as well as the 45-degree line of random classification.



(a) t-stat distribution

(b) Receiver Operating Characteristics (ROC)

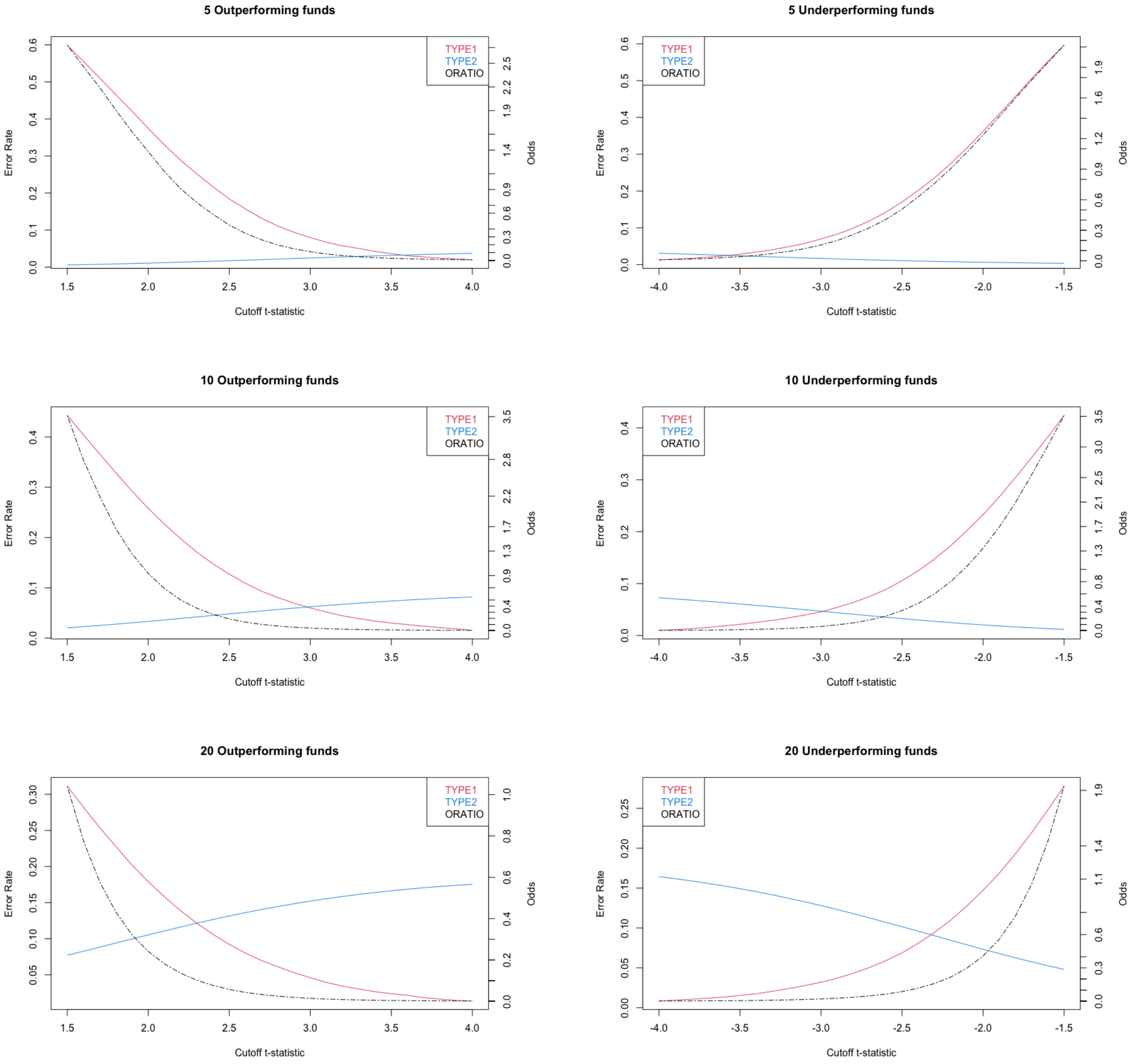
We first had to determine a threshold for performance evaluation that produces a specific false discovery rate to analyze the error rates. As Harvey & Liu (2020), we control at a 5 percent False Discovery Rate in our analyses of the cutoff t-statistics. Next, we create a vector of parameters for the Type II error rate to correspond to the alternative hypotheses currently in testing. These parameters are denoted as p_0, and the vector contains values of 0, 5, 10 and 20. The parameter p_0 represent the fraction of managers initially assumed to have skill. In turn, the assumption of p_0 managers having skill implies that $1 - p_0$ of managers does not have skill. By adjusting the sample regarding these assumptions, we create a matrix, where we define p_0 of funds to have the skill and contain their respective alpha-values. The remaining sample, $1 - p_0$, is assumed to have no skill, and thus the alpha-values of these funds are set to zero, implying that their return is fully explained by the factors of the Carhart (1997) four-factor model. We will also assume that a fraction of managers are unskilled, implying that p_0 is unskilled. Thus, we range the matrix with the p_0 worst funds containing their alpha to test the absence of skill.

Calculation of error rates Type I, Type II, and the Oratio is described in detail in Section 3.3.2. We receive a distribution dependant upon the cutoff t-statistic for each p_0; 5, 10 and 20 out- and underperforming funds through the formulas and the computation of the error rates.

These results are displayed in Figure 11 and in Appendix B. We will now work with a new set of hypotheses to fit the new performance evaluation methodology as described in Section 3.3.2.

Figure 11: Error rates for fixed t-statistic thresholds

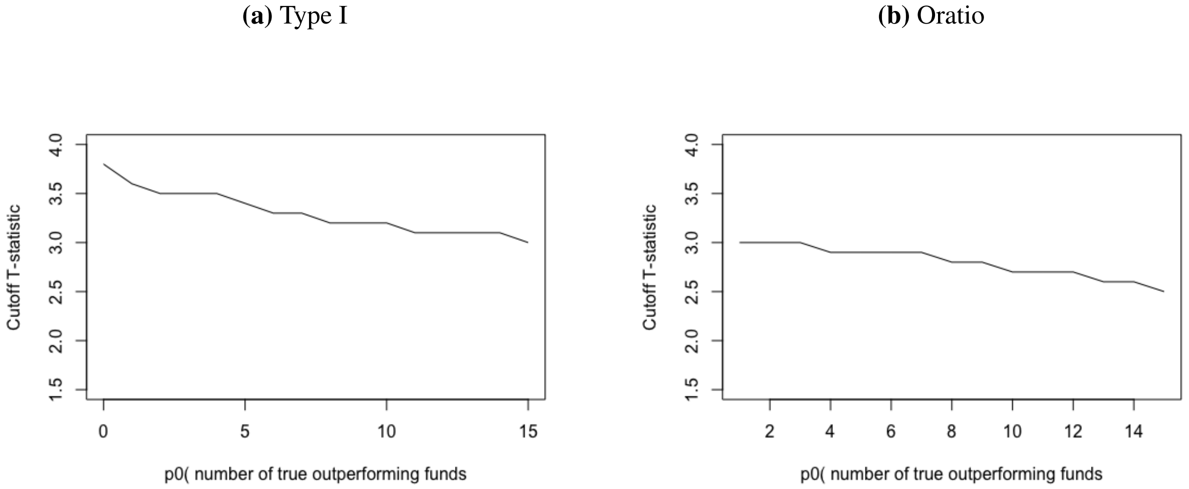
The figure shows the simulated distribution of error rates for values of $p_0 = 5, 10, 20$ out- and underperforming funds. The graphs display the development of Type I and Type II error rates and Oratio, varying on the cutoff t-statistic increments. The figures are simulated from 107 fund's excess return and bootstrapped 1000x100 times. The red line represents the Type I error rate, the blue line represents the Type II error rate, and the black dotted line represents Oratio. The left axis is the error rate of Type I and II. The right axis is the Oratio values. The x-axis is the optimal cutoff t-statistics increment from (1.5 to 4.0) or (-4.0 to -1.5).



In Figure 11, for all p_0, we see that the Type I error rate is declining when the cutoff t-statistic increases and the Type II error rate increases. In this demonstration, we present the Type II error rate that is proposed by Harvey & Liu (2020), not the standard approach. Type I error rate is high when $p_0 = 5$, starting at 0.6 at the 1.5 cutoff t-statistic, while Type II is at its lowest, starting at 0.006. The error rates have a clear tendency towards a tradeoff. With an increase in cutoff t-statistic, we see the error rates trade off on one another. The Oratio represents this relationship drawing a curve closely following the Type I error rate but still affected downwards by the Type II error rate. The higher the demand of the t-statistical cutoff, the lower the possibility of performing a Type I error rate. This is mainly because high alpha t-statistics provide less chance of identifying false positives out of the number of true positives. Type II error rate increases as we increase the demand for t-statistics. The higher the demand, the more likely to define a manager as unskillful when he is not. We see this more clearly when increasing p_0 to 20. The tradeoff is more apparent as the Type II ratio relative to Type I increases for all t-statistics. When we assume more skilled managers, the possibility of rejecting a true positive increase, while the ratio for accepting false positives decreases. Due to the cutoff analysis, the procedure is robust as it considers which cutoff values for these funds correspond to the 5% false discovery rate. As we increase our prior on p_0, the error rates differ from the lower priors on p_0, as should be expected.

Figure 12: Optimal error rate cutoff t-statistics

The figures display optimal error rate cutoffs for various p_0 for the Type I and Oratio. Figure a) display the optimal cutoff computed by controlling for 5 percent FDR from the thresholds distribution in Figure 11 and Table 11. Figure b) displays the optimal t-statistical cutoff for the Oratio while controlling for the relationship between Type I and Type II at 0.1, when controlling for the assumption of Type I being ten times as costly as Type II.


(a) Type I

(b) Oratio

If an investor would evaluate performance based upon the framework presented, we provide a measurement tool through Oratio. Figure 11 and Appendix B or C may be used as a function of cost for a potential investor. Suppose the investor believes that a false positive is ten times more costly than a false negative. The investor could use Appendix B or C, and use 1/10 as a benchmark for the Oratio, weighting the error rates according to cost, and thus determine optimal cutoff t-statistic for his specific cost assessment. Figure 12 display the optimal cutoff t-statistic for various H_A. Figure a) display the cutoff t-statistic when controlling for 5% Type I error rate. Figure b) display the cutoff t-statistic when Oratio is 0.1, as illustrated in the example above.

Table 10: Double bootstrap results

The table displays the summary of optimal cutoffs on a range of p_0 from the results of the second bootstrap of Harvey & Liu (2020). It displays the new optimal cutoff t-statistics for the p_0. By using the optimal cutoff as a reference, it is measured towards the ranked t-statistic of each funds alpha, and thus determine whether the H_0 is rejected or not. Table 10 further display the number of observations with a t-statistic, higher for outperforming or lower for underperforming, than the respective cutoff of the particular p_0, or H_0 in question.

p_0	*0*	*5*	*10*	*20*
Outperforming funds				
Optimal cutoff	4.0<	3.3	3.2	3.0
No. of funds outperf.	0	0	0	0
H_0 rejected?	No	**Yes**	**Yes**	**Yes**
Test power (1-Type II)	0	55.8%	58.1%	68.1%
Underperforming funds				
Optimal cutoff	-4.0>	-3.2	-3.0	-2.8
No. of funds underperf.	0	1	2	3
H_0 rejected?	**Yes**	No	No	No
Test power (1-Type II)	0	74.6%	79.6%	83,4%

Table 10 displays whether there exists skill among mutual funds with the assumption that a fraction of managers have are outperforming or underperforming. We find the optimal cutoff for a 5 percent Type I error rate that we further use to estimate the Type II error. In this section, we use the standard approach to estimate Type II errors, not the method suggested by Harvey & Liu (2020). We change to S-Type II error rate as the standard approach allow us to calculate the Test power. We can reject the null hypothesis if none of the funds exhibit a higher t-statistic than the cutoff value, indicating no skillful managers. If any funds exhibit an alpha t-statistic higher than the cutoff, this implies that we do not reject the null hypothesis and proceed to the alternative hypothesis. If the test has a low Type II error (false negative), the Test power is high, indicating that we have a majority of true positives. There is no formal significance standard for Test power to be adequate. However, an 80% Test power should be sufficient to say, with certainty, that there exists outperforming or underperforming managers as a result of skill or lack thereof.

From Table 10, we further aim to locate whether or not our assumptions on p_0 out- or underperforming funds are true or not. The table display the results for $p_0 = 0, 5, 10, 20$. The cutoff t-statistic is higher for p_0 in the second bootstrap than in the first bootstrap, as seen in Section 5.4.1 - Cutoff t-statistics. The number of observed funds above the optimal cutoff t-statistic by controlling for $FDR = 0.05$ is zero for all observations for outperforming funds. We reject the H_0 for every p_0, and we thus conclude that $p_0 = 5, 10, 20$ number of funds are not outperforming their benchmark with statistical significance as a result of skill. However, the Type II error rates display the ratio of misses in the sample. For higher values of p_0, the ratio decreases. The error rates is presented in Appendix B and C. In essence, we can say that by controlling for 0.05 FDR, with a cutoff of 3.0 with the assumption of 20 outperforming funds, the rate of missed outperforming managers are 0.319. We receive a Test power of 68.1%. A Test power of 80% (probability of Type II = 20%) is necessary to ensure the probability of the error rate is sufficiently low in the test. Even though we reject H_0 with a Test power of 68.1%, we cannot find skillful managers. However, we cannot say for certain that no skillful managers exist in our sample. We illustrate how the error rates are estimated for $p_0 = 20$ in Figure 13.

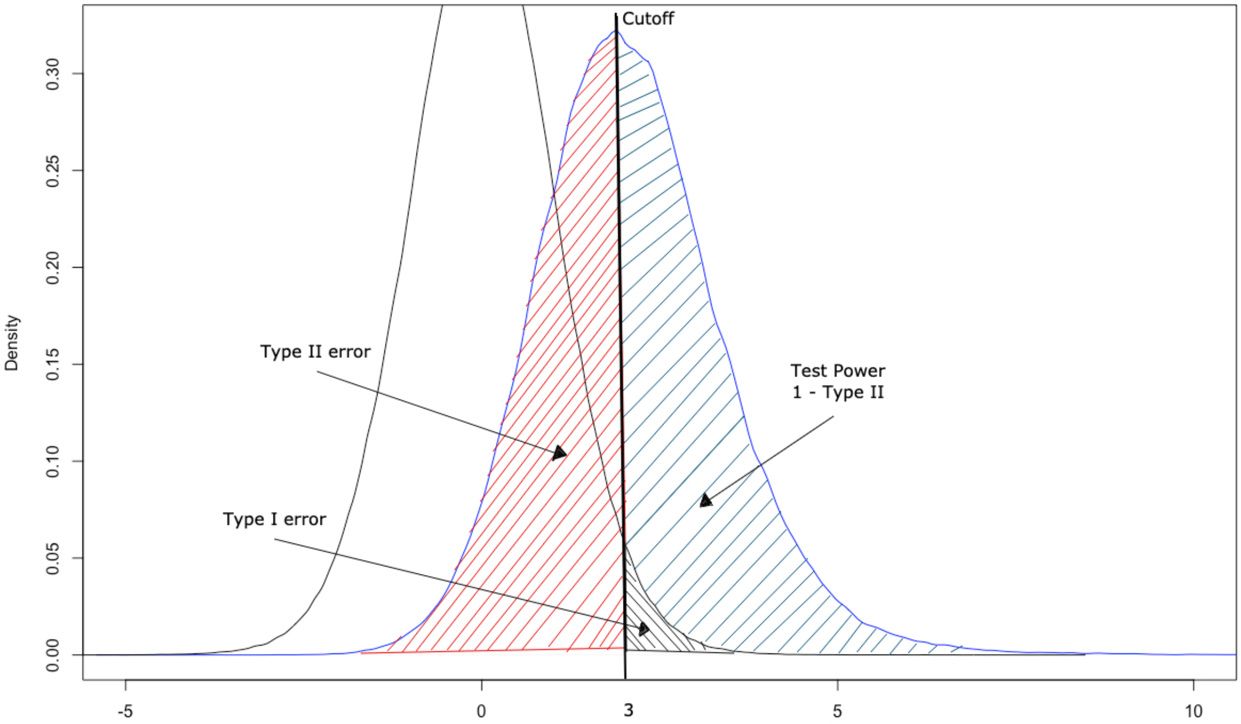
For the underperforming funds, we receive different results. By analyzing the cutoff for different p_0, we identify some funds conveying a t-statistic lower than the negative cutoff threshold. The result implies that we are not able to reject H_0. For $p_0 = 5$, we locate one underperforming fund with statistical significance. For $p_0 = 10$ we have two funds and for $p_0 = 20$, we have three funds. The discovery of funds with higher t-statistics than the thresholds tells us that the H_0 can not be rejected as presented above. This result further indicates that there is a possibility that there, in fact, exists underperforming funds. However, we cannot reject the H_0. We also see that the Type II ratio decreases for an increasing number of p_0, implying that the Type II error probability decreases, and we receive a higher Test power. The Test power when assuming $p_0 = 20$ is above the required threshold of 80% (83.4%). Moreover, we can conclude with high certainty that underperforming Norwegian mutual fund managers exist due to an absence of skills.

We will now illustrate how optimal cutoff for Oratio will affect our test of the null hypothesis. Harvey & Liu (2020) suggests several advantages towards the use of Oratio as it quantifies the chance of a false discovery per miss. In our example, we illustrate the Oratio when we suppose that the cost of a Type I error is ten times that of a Type II error. We can use Appendix B to control for the first cutoff value below 0.1. As we illustrated in Figure 12, the optimal cutoff for Oratio is not very different from Type I under the assumption of outperforming managers. However, when we test the null hypotheses using the cutoff t-statistic for Oratio for underperforming managers from Appendix C, we get another slightly different result. For $p_0 = 10$ there are two underperforming funds and for $p_0 = 20$ we identify four underperforming funds. However, the main conclusion remains the same.

Figure 13 illustrate the error rates under the assumption of $p_0 = 20$ managers are outperforming. The blue distribution is where the density of t-statistics under the assumption of $\alpha \neq 0$ and the black distribution is under the density of t-statistics under the assumption that $\alpha = 0$. In this case, we do not observe any t-statistics above the cutoff threshold, indicating that we reject the null hypothesis. However, there is a large possibility of Type II error (red area), that we miss a manager we taught was unskilled. We cannot, with certainty, say that skillful managers do not exist in our sample.

Figure 13: Error rate illustration

The figure displays the double bootstrapped results under the assumption of zero performance to the left (black distribution) and under the assumption that p_0 managers have skills. The bold line present the cutoff t-statistic, the marked red area is the probability of Type II error, the marked black area is the probability of a Type I error, and the marked blue area is the Test power. The illustration is presented under the assumption of $p_0 = 20$ with an FDR = 5%



6 Conclusion

Our thesis evaluate the performance of mutual fund managers using data from 1987 to 2019 on 107 Norwegian mutual funds. The primary performance model we use in our tests is the Carhart (1997) four-factor model. First, we test managers' performance on an aggregate level. The result indicates that the mutual fund managers produce a nonsignificant alpha at 0.035, suggesting that the managers cannot beat the benchmark beyond their management cost on an aggregate level. Second, we employ the individual standard test to find the actual performance of each manager and their parametric p-value. The generated alpha values imply that some managers can beat the benchmark beyond their fees and that some fail to beat the benchmark.

Further, we use a bootstrap approach to distinguish if the generated alpha results from luck or skill. We follow a similar approach to Kosowski et al. (2006) and Fama & French (2010) to test the null hypothesis of zero performance for all funds, both individually and jointly. Through the bootstrapped p-value, we fail to identify skill among the best performing funds, as we identify that their performance results from luck rather than skill. However, for the worst-performing funds, we reject the null hypothesis. We identify significant bootstrap p-values that indicate that their returns result from bad performances rather than unlucky performances.

To present novelty to the subject, we implement the double bootstrap procedure similar to Harvey & Liu (2020). This procedure allowed us to estimate the Type II error under the null hypothesis that a fraction of managers outperform. We follow their procedure to estimate the cutoff t-statistic by controlling for a certain level of Type I error rate in order to characterize the Type II error rate. We find no statistically significant evidence that skillful managers exist. We reject the null hypothesis: a specific fraction of managers outperform. However, our estimated Test power of 50% indicates that there could be skilled outperforming managers, though we cannot detect these. When we test the null hypothesis that a fraction of managers is underperforming, we fail to reject the null hypothesis. We find statistically significant evidence that some managers of Norwegian mutual funds display a lack of skill.

To summarize, we find significant evidence that there exist unskilled managers. We also find that some Norwegian mutual fund managers generate positive abnormal returns, but these returns result from luck. However, the estimated Test power indicates that some skillful managers might exist, but our tests are not powerful enough to detect them.

7 References

[Bjerke and Bråthe, 2020] Bjerke, E. A. and Bråthe, N. (2020). The paradox of skill in norwegian mutual funds. Master's thesis, University of Agder.

[Brown et al., 1992] Brown, S. J., Goetzmann, W., Ibbotson, R. G., and Ross, S. A. (1992). Survivorship Bias in Performance Studies. *The Review of Financial Studies*, 5(4).

[Busse et al., 2010] Busse, J. A., Goyal, A., and Wahal, S. (2010). Performance and persistence in institutional investment management. *The Journal of Finance*, 65(2).

[Carhart, 1997] Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1).

[Daniel et al., 1997] Daniel, K., Grinblatt, M., Titman, S., and Wermers, R. (1997). Measuring Mutual Fund Performance with Characteristic-Based Benchmarks. *The Journal of Finance*, 52(3).

[Elton et al., 1996] Elton, E. J., Gruber, M. J., and Blake, C. R. (1996). The persistence of risk-adjusted mutual fund performance. *The Journal of Business*, 69(2).

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

[Fama and French, 2010] Fama, E. F. and French, K. R. (2010). Luck versus Skill in the cross-section of mutual fund returns. *The Journal of Finance*, 65(5).

[Grinblatt and Titman, 1992] Grinblatt, M. and Titman, S. (1992). The Persistence of Mutual Fund Performance. *The Journal of Finance*, 47(5).

[Gruber, 1996] Gruber, M. J. (1996). Another puzzle: The growth in actively managed mutual funds. *The Journal of Finance*, 51(3).

[Harvey and Liu, 2020] Harvey, C. R. and Liu, Y. (2020). False (and Missed) Discoveries in Financial Economics. *The Journal of Finance*, 75(5).

[Harvey et al., 2016] Harvey, C. R., Liu, Y., and Zhu, H. (2016). and the Cross-Section of Expected Returns. *The Review of Financial Studies*, 29(1).

[Jegadeesh and Titman, 1993] Jegadeesh, N. and Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, 48(1).

[Jensen, 1968] Jensen, M. C. (1968). The performance of mutual funds in the period 1945-1964. *The Journal of Finance*, 23(2).

[Kosowski et al., 2006] Kosowski, R., Timmermann, A., Wermers, R., and White, H. (2006). Can mutual fund ”stars” really pick stocks? New evidence from a bootstrap analysis. *The Journal of Finance*, 61(6).

[Lintner, 1965] Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance*, 20(4).

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

[Sørensen, 2011] Sørensen, L. Q. (2011). Mutual Fund Performance at the Oslo Stock Exchange. *The SSRN Electronic Journal*.

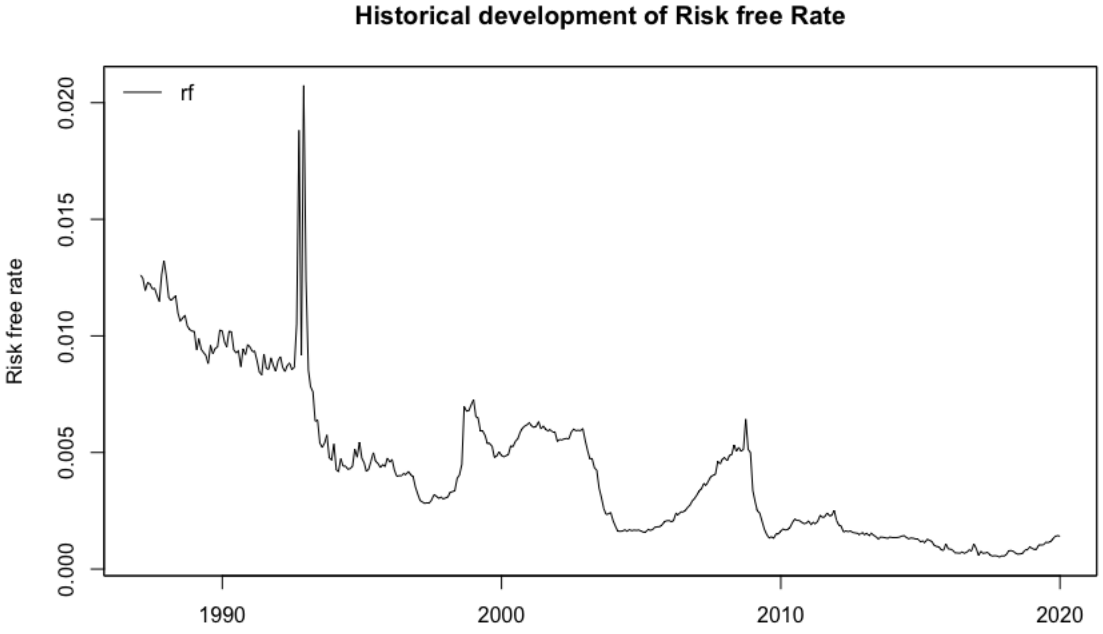
[Ødegaard, 2011] Ødegaard, B. A. (2011). Empirics of the Oslo Stock Exchange. Basic, descriptive, results. *University of Stavanger*.

[Ødegaard, 2019] Ødegaard, B. A. (2019). Asset pricing data at OSE market returns. `https://ba-odegaard.no/financial_data/ose_asset_pricing_data/index.html`.

8 Appendix

Appendix A: Historical risk free rate development

The figure displays the monthly development of the Norwegian risk free rate (NIBOR) for the full sample period from 1987 to 2019.



Appendix B: Error rates for fixed t-statistics for outperforming funds

The table present the findings of the Harvey & Liu (2020) double bootstrap method with the estimation of Type I, Type II and Oratio for outperforming funds in our sample of Norwegian mutual funds. The left side of the table represent the hurdle of t-statistic cutoffs ranging from 1.5 to 4.0. The first row represent the different priors on p_0 in the analyses: 0, 5, 10 and 20 funds outperforming, respectively. The numbers displayed in the table represent the rate of errors for a given prior on p_0 and a given cutoff t-statistic from the double bootstrapped sample of fund returns.

t-stat	Type I				Type II			Oratio		
	0	5	10	20	5	10	20	5	10	20
1,5	1.000	0.599	0.443	0.311	0.061	0.063	0.045	2.730	3.520	1.042
1,6	1.000	0.555	0.404	0.282	0.072	0.073	0.054	2.458	2.788	0.777
1,7	1.000	0.510	0.366	0.254	0.085	0.085	0.065	2.194	2.198	0.578
1,8	0.999	0.466	0.329	0.228	0.099	0.097	0.076	1.907	1.667	0.432
1,9	0.997	0.420	0.292	0.202	0.112	0.112	0.089	1.631	1.254	0.321
2,0	0.993	0.374	0.258	0.179	0.128	0.130	0.103	1.381	0.935	0.241
2,1	0.987	0.330	0.227	0.158	0.146	0.149	0.119	1.127	0.686	0.181
2,2	0.976	0.288	0.198	0.139	0.165	0.170	0.137	0.910	0.501	0.136
2,3	0.958	0.251	0.171	0.121	0.184	0.189	0.154	0.734	0.365	0.102
2,4	0.938	0.217	0.148	0.106	0.203	0.209	0.174	0.585	0.265	0.077
2,5	0.912	0.184	0.127	0.092	0.227	0.233	0.194	0.449	0.190	0.057
2,6	0.879	0.157	0.109	0.080	0.252	0.258	0.219	0.345	0.137	0.043
2,7	0.844	0.132	0.093	0.070	0.275	0.282	0.242	0.262	0.101	0.033
2,8	0.809	0.111	0.080	0.061	0.297	0.305	0.267	0.196	0.074	0.025
2,9	0.774	0.094	0.069	0.053	0.327	0.332	0.292	0.148	0.054	0.019
3,0	0.732	0.080	0.060	0.046	0.355	0.362	0.319	0.112	0.041	0.014
3,1	0.692	0.067	0.052	0.039	0.380	0.391	0.345	0.082	0.030	0.011
3,2	0.654	0.057	0.044	0.034	0.411	0.419	0.371	0.062	0.023	0.008
3,3	0.618	0.050	0.039	0.030	0.442	0.447	0.401	0.048	0.018	0.006
3,4	0.582	0.042	0.034	0.027	0.470	0.474	0.427	0.036	0.013	0.005
3,5	0.547	0.036	0.030	0.024	0.496	0.502	0.455	0.028	0.011	0.004
3,6	0.510	0.030	0.027	0.022	0.521	0.529	0.481	0.022	0.008	0.003
3,7	0.478	0.027	0.024	0.018	0.551	0.557	0.509	0.017	0.007	0.003
3,8	0.446	0.024	0.021	0.016	0.580	0.582	0.532	0.013	0.005	0.002
3,9	0.416	0.022	0.018	0.015	0.607	0.608	0.559	0.011	0.004	0.002
4,0	0.384	0.020	0.015	0.014	0.633	0.634	0.585	0.009	0.003	0.002

Appendix C: Error rates for fixed t-statistics for underperfoming funds

The table present the findings of the Harvey & Liu (2020) double bootstrap method with the estimation of Type I, Type II and Oratio for underperforming funds in our sample of Norwegian mutual funds. The left side of the table represent the hurdle of t-statistic cutoffs ranging from -4.0 to -1.5. The first row represent the different priors on p_0 in the analyses: 5, 10 and 20 funds underperforming, respectively. The numbers displayed in the table represent the rate of errors for a given prior on p_0 and a given cutoff t-statistic from the double bootstrapped sample of fund returns.

t-stat		Type I			Type II			Oratio		
	0	5	10	20	5	10	20	5	10	20
-4,0		0.013	0.010	0.008	0.458	0.442	0.444	0.010	0.004	0.002
-3,9		0.015	0.011	0.009	0.429	0.417	0.415	0.014	0.005	0.002
-3,8		0.017	0.013	0.011	0.404	0.390	0.386	0.018	0.006	0.003
-3,7		0.020	0.015	0.012	0.377	0.362	0.360	0.023	0.008	0.003
-3,6		0.024	0.018	0.013	0.352	0.335	0.334	0.032	0.011	0.004
-3,5		0.028	0.021	0.016	0.326	0.312	0.309	0.041	0.015	0.005
-3,4		0.034	0.025	0.018	0.301	0.287	0.285	0.054	0.020	0.006
-3,3		0.040	0.029	0.021	0.277	0.263	0.262	0.071	0.027	0.009
-3,2		0.049	0.034	0.024	0.254	0.243	0.238	0.093	0.036	0.011
-3,1		0.058	0.039	0.028	0.235	0.223	0.218	0.120	0.049	0.015
-3,0		0.070	0.045	0.032	0.213	0.204	0.201	0.156	0.067	0.020
-2,9		0.083	0.054	0.037	0.196	0.187	0.183	0.199	0.094	0.027
-2,8		0.099	0.064	0.043	0.180	0.170	0.166	0.259	0.127	0.035
-2,7		0.119	0.075	0.051	0.164	0.154	0.150	0.324	0.173	0.047
-2,6		0.142	0.088	0.059	0.148	0.141	0.137	0.407	0.236	0.064
-2,5		0.170	0.106	0.069	0.136	0.125	0.126	0.507	0.327	0.086
-2,4		0.202	0.125	0.081	0.123	0.112	0.115	0.628	0.444	0.117
-2,3		0.236	0.147	0.094	0.111	0.102	0.106	0.760	0.599	0.159
-2,2		0.275	0.173	0.110	0.099	0.093	0.098	0.904	0.799	0.217
-2,1		0.317	0.202	0.128	0.089	0.084	0.090	1.066	1.049	0.298
-2,0		0.362	0.233	0.147	0.081	0.075	0.082	1.235	1.345	0.407
-1,9		0.409	0.267	0.169	0.074	0.068	0.074	1.416	1.692	0.558
-1,8		0.457	0.304	0.193	0.066	0.060	0.065	1.597	2.092	0.768
-1,7		0.505	0.343	0.220	0.060	0.054	0.059	1.775	2.548	1.057
-1,6		0.552	0.383	0.248	0.054	0.048	0.054	1.947	3.023	1.442
-1,5		0.596	0.424	0.278	0.049	0.042	0.048	2.118	3.515	1.939

Appendix D: Descriptive statistics Norwegian mutual fund

1

Name	Obs	Mean	Std.dev	Kurt	Skew	Max	Min
ABIF Norge ++	56.00	0.01	0.07	-0.47	-0.31	0.14	-0.16
Alfred Berg Aksjef Norge	115.00	0.01	0.06	1.88	-0.77	0.13	-0.25
Alfred Berg Aksjespar	106.00	0.01	0.07	2.17	-0.87	0.13	-0.28
Alfred Berg Aktiv	289.00	0.01	0.07	2.59	-0.77	0.21	-0.27
Alfred Berg Aktiv II	182.00	0.01	0.07	1.23	-0.60	0.18	-0.27
Alfred Berg Gambak	350.00	0.01	0.07	2.69	-0.40	0.28	-0.27
Alfred Berg Humanfond	241.00	0.01	0.06	3.04	-0.97	0.16	-0.26
Alfred Berg N. Pensjon	52.00	0.01	0.06	4.56	-1.36	0.12	-0.25
Alfred Berg Norge	147.00	0.01	0.07	1.90	-0.96	0.17	-0.27
Alfred Berg Norge gml	197.00	0.01	0.07	2.23	-0.97	0.17	-0.27
Alfred Berg Norge Classic	351.00	0.01	0.06	3.04	-1.06	0.17	-0.27
Alfred Berg Norge Etisk	146.00	0.01	0.07	2.58	-1.04	0.17	-0.28
Alfred Berg Norge Inst	72.00	0.01	0.03	1.26	-0.95	0.07	-0.08
Alfred Berg Vekst	72.00	0.01	0.08	1.89	-0.51	0.19	-0.28
Arctic Norwegian Equities Class A	109.00	0.01	0.03	1.51	-0.67	0.09	-0.09
Arctic Norwegian Equities Class B	110.00	0.01	0.03	1.44	-0.57	0.10	-0.09
Arctic Norwegian Equities Class D	83.00	0.01	0.03	1.55	-0.92	0.07	-0.09
Arctic Norwegian Equities Class I	110.00	0.01	0.03	1.41	-0.58	0.10	-0.09
Atlas Norge	263.00	0.01	0.07	3.67	-0.09	0.37	-0.25
Banco Norge	38.00	0.01	0.07	-0.27	-0.33	0.14	-0.17
C WorldWide Norge	294.00	0.01	0.06	3.07	-0.89	0.20	-0.28
Carnegie Aksje Norge	210.00	0.01	0.07	2.06	-0.86	0.20	-0.28
Danske Invest Aktiv Formuesf. A	21.00	0.01	0.04	0.59	-0.83	0.08	-0.11
Danske Invest Norge Aksj. Inst 1	237.00	0.01	0.06	2.69	-0.93	0.15	-0.23
Danske Invest Norge Aksj. Inst 2	158.00	0.01	0.05	4.29	-1.16	0.15	-0.23
Danske Invest Norge I	312.00	0.01	0.06	3.63	-1.03	0.15	-0.29
Danske Invest Norge II	312.00	0.01	0.06	3.64	-1.02	0.15	-0.29
Danske Invest Norge Vekst	312.00	0.01	0.06	6.61	0.33	0.42	-0.26
Delphi Norge	307.00	0.01	0.07	2.07	-0.54	0.23	-0.25
Delphi Vekst	193.00	0.01	0.08	1.03	-0.33	0.26	-0.23
DNB Norge	289.00	0.01	0.06	2.38	-0.84	0.16	-0.24
DNB Norge (Avanse I)	327.00	0.01	0.06	2.10	-0.96	0.16	-0.26
DNB Norge (Avanse II)	287.00	0.01	0.06	2.39	-0.96	0.16	-0.26
DNB Norge (I)	295.00	0.01	0.07	15.28	1.31	0.59	-0.24
DNB Norge (III)	283.00	0.01	0.06	2.37	-0.87	0.16	-0.24
DNB Norge (IV)	206.00	0.01	0.06	2.95	-0.89	0.16	-0.24
DNB Norge R	12.00	0.01	0.03	0.70	-1.20	0.04	-0.06
DNB Norge Selektiv (II)	214.00	0.01	0.06	2.37	-0.77	0.17	-0.24

Continued on next page

Table 11 – *Continued from previous page*

Name	Obs	Mean	Std.dev	Kurt	Skew	Max	Min
DNB Norge Selektiv (III)	307.00	0.01	0.06	2.22	-0.82	0.17	-0.24
DnB Real-Vekst	157.00	0.01	0.09	24.68	2.15	0.69	-0.40
DNB SMB	226.00	0.01	0.07	1.17	-0.47	0.17	-0.26
Eika Norge	196.00	0.01	0.06	3.96	-1.02	0.18	-0.25
Eika SMB	187.00	0.01	0.07	1.28	-0.67	0.17	-0.23
FIRST Generator	112.00	0.01	0.06	1.41	-0.77	0.16	-0.19
FIRST Norge Fokus	14.00	0.01	0.03	0.70	-0.99	0.05	-0.06
Fokus Barnespar	32.00	-0.00	0.08	3.44	-1.21	0.13	-0.28
Fondsfinans Aktiv II	48.00	-0.00	0.07	-0.07	-0.23	0.14	-0.16
Fondsfinans Norge	205.00	0.01	0.06	2.66	-0.78	0.16	-0.26
FORTE Norge	107.00	0.01	0.04	1.05	-0.08	0.14	-0.12
FORTE Tr?nder	81.00	0.01	0.03	0.27	-0.14	0.09	-0.09
GAMBAK Oppkj?p	19.00	0.00	0.05	0.48	0.35	0.14	-0.09
GJENSIDIGE AksjeSpar	152.00	0.01	0.07	2.18	-0.94	0.17	-0.27
GJENSIDIGE Invest	104.00	0.01	0.06	2.39	-0.85	0.13	-0.21
Globus Aktiv	88.00	0.01	0.08	0.35	-0.30	0.24	-0.23
Globus Norge	103.00	0.01	0.08	0.38	-0.35	0.22	-0.23
Globus Norge II	95.00	0.01	0.08	0.40	-0.24	0.23	-0.23
Handelsbanken Norge	300.00	0.01	0.06	4.08	-1.17	0.18	-0.29
Handelsbanken Norge A10	18.00	0.00	0.04	0.52	-1.18	0.05	-0.09
Holberg Norge	229.00	0.01	0.06	1.70	-0.52	0.16	-0.24
K-IPA Aksjefond	37.00	0.01	0.07	2.08	-0.97	0.12	-0.22
KLP Aksjeinvest	97.00	0.00	0.06	1.67	-0.78	0.15	-0.22
KLP AksjeNorge	250.00	0.01	0.06	3.24	-0.91	0.18	-0.30
Landkreditt Norge	122.00	0.01	0.06	2.14	-0.74	0.17	-0.21
Landkreditt Utbytte	83.00	0.01	0.02	0.44	-0.76	0.05	-0.05
Landkreditt Utbytte I	19.00	0.01	0.02	-0.29	-0.39	0.04	-0.04
NB-Aksjefond	207.00	0.01	0.06	2.14	-0.94	0.18	-0.25
Nordea Avkastning	396.00	0.01	0.06	2.70	-0.87	0.21	-0.28
Nordea Barnespar	47.00	-0.00	0.06	-0.32	-0.36	0.11	-0.16
Nordea Kapital	298.00	0.01	0.06	2.92	-1.01	0.17	-0.26
Nordea Kapital II	84.00	0.01	0.07	-0.20	-0.47	0.13	-0.17
Nordea Kapital III	70.00	0.01	0.07	-0.25	-0.56	0.13	-0.17
Nordea Norge Pluss	105.00	0.01	0.04	1.17	-0.65	0.12	-0.11
Nordea Norge Verdi	287.00	0.01	0.05	2.66	-0.87	0.15	-0.24
Nordea SMB	213.00	0.01	0.07	0.54	-0.23	0.18	-0.23
Nordea SMB II	70.00	-0.01	0.08	0.13	0.17	0.19	-0.19
Nordea Vekst	337.00	0.01	0.07	1.85	-0.85	0.20	-0.26
ODIN Norge	331.00	0.01	0.06	2.42	-0.43	0.23	-0.24
ODIN Norge A	50.00	0.01	0.03	1.55	-1.19	0.05	-0.09

Continued on next page

Table 11 – *Continued from previous page*

Name	Obs	Mean	Std.dev	Kurt	Skew	Max	Min
ODIN Norge B	50.00	0.01	0.03	1.56	-1.20	0.05	-0.09
ODIN Norge D	50.00	0.01	0.03	1.56	-1.20	0.05	-0.09
ODIN Norge II	139.00	0.01	0.06	3.11	-0.99	0.14	-0.24
Orkla Finans 30	162.00	0.02	0.06	1.48	-0.71	0.15	-0.26
Pareto Aksje Norge	220.00	0.01	0.05	3.53	-0.84	0.16	-0.26
PLUSS Aksje (Fondsforval)	277.00	0.01	0.06	2.32	-0.72	0.18	-0.26
PLUSS Markedsverdi (Fondsforv)	300.00	0.01	0.06	3.21	-0.98	0.16	-0.25
Postbanken Aksjevekst	97.00	0.01	0.07	0.15	-0.40	0.15	-0.20
RF Aksjefond	116.00	0.01	0.06	1.27	-0.73	0.14	-0.24
RF-Plussfond	54.00	0.01	0.07	-0.53	-0.36	0.14	-0.17
Sbanken Framgang Sammen	47.00	0.01	0.03	0.81	-0.73	0.07	-0.07
SEB Norge LU	67.00	-0.00	0.07	1.24	-0.65	0.16	-0.26
Skandia Horisont	97.00	0.01	0.06	1.13	-0.76	0.16	-0.22
Skandia SMB Norge	97.00	0.00	0.07	2.44	-1.01	0.14	-0.27
SR-Bank Norge A	12.00	0.01	0.03	-0.50	-0.49	0.05	-0.05
SR-Bank Norge B	12.00	0.01	0.03	-0.50	-0.49	0.05	-0.05
Storebrand Aksje Innland	282.00	0.01	0.06	3.08	-1.02	0.15	-0.27
Storebrand AksjeSpar	226.00	0.01	0.04	1.30	-0.89	0.10	-0.14
Storebrand Norge	396.00	0.01	0.06	2.48	-0.89	0.17	-0.29
Storebrand Norge A	43.00	0.02	0.07	-0.24	-0.52	0.15	-0.17
Storebrand Norge Fossilfri	33.00	0.01	0.02	1.35	-0.89	0.05	-0.05
Storebrand Norge I	237.00	0.01	0.06	3.21	-1.00	0.15	-0.29
Storebrand Norge Institusjon	39.00	0.01	0.04	0.65	-0.53	0.10	-0.10
Storebrand Optima Norge	221.00	0.01	0.06	3.04	-0.99	0.15	-0.29
Storebrand Vekst	328.00	0.01	0.07	3.70	0.01	0.37	-0.30
Storebrand Verdi	265.00	0.01	0.06	3.18	-0.97	0.14	-0.27
Storebrand Verdi N	22.00	0.01	0.03	-0.15	-0.53	0.06	-0.06
Terra Norge	187.00	0.01	0.07	1.46	-0.75	0.19	-0.26
VAAR Aksjefond	39.00	0.01	0.07	3.49	-1.19	0.11	-0.26

[1] The table present discriptive statistics for all 107 mutual funds in our sample. The funds are ranked alphabetically. The descriptive statistics represent the following. Column 2 show the number of observations, e.g. the number of months of historical data. Column 3 provide the excess return of the fund, net of management fees. Columns 4 till 8 show standard statistical measurements of the sample: standard deviation, kurtosis, skew, maximum observed value and minimum observed value for each fund, respectively.

Appendix E: Bootstrap results for individual mutual funds

2

Name	Alpha	t-stat alpha	p-value	boot p-value
Danske Invest Norge Aksj. Inst 1	3.02	2.23	0.01	0.80
FIRST Norge Fokus	10.95	2.19	0.01	0.50
Landkreditt Utbytte	5.38	2.15	0.02	0.35
Fondsfinans Norge	3.78	1.94	0.03	0.37
Danske Invest Norge Aksj. Inst 2	3.26	1.94	0.03	0.21
PLUSS Markedsverdi (Fondsforv)	1.95	1.83	0.03	0.21
Landkreditt Utbytte I	7.62	1.82	0.03	0.10
FORTE Trxb0nder	6.18	1.43	0.08	0.66
Storebrand Norge I	1.60	1.36	0.09	0.64
Storebrand Norge Fossilfri	4.18	1.24	0.11	0.78
Storebrand Optima Norge	1.72	1.19	0.12	0.77
Landkreditt Norge	3.19	1.15	0.12	0.69
Carnegie Aksje Norge	1.78	1.15	0.13	0.62
DNB Norge R	6.11	1.11	0.13	0.55
Danske Invest Norge II	1.39	1.07	0.14	0.60
Storebrand Verdi N	2.32	1.02	0.15	0.58
Storebrand Verdi	1.13	0.99	0.16	0.57
Eika Norge	1.71	0.97	0.17	0.54
Pareto Aksje Norge	1.80	0.95	0.17	0.43
K-IPA Aksjefond	4.72	0.90	0.18	0.50
Nordea Norge Verdi	1.35	0.88	0.19	0.45
Alfred Berg Norge	1.37	0.85	0.20	0.43
PLUSS Aksje (Fondsforval)	1.25	0.82	0.21	0.43
Nordea Kapital	0.93	0.81	0.21	0.40
Storebrand Norge	0.97	0.74	0.23	0.53
C WorldWide Norge	0.92	0.74	0.23	0.44
DNB SMB	1.91	0.71	0.24	0.44
Alfred Berg Norge Inst	1.30	0.66	0.26	0.54
Holberg Norge	1.15	0.59	0.28	0.61
ODIN Norge	1.07	0.55	0.29	0.65
Danske Invest Norge I	0.72	0.55	0.29	0.58
ODIN Norge A	1.55	0.54	0.29	0.52
DNB Norge Selektiv (II)	0.58	0.50	0.31	0.55
ODIN Norge D	1.31	0.45	0.32	0.59
ODIN Norge B	1.28	0.45	0.33	0.58
Vxc5R Aksjefond	1.91	0.42	0.34	0.58
Storebrand Aksje Innland	0.23	0.30	0.38	0.80
Nordea Avkastning	0.48	0.29	0.38	0.75

Continued on next page

Table 12 – *Continued from previous page*

Name	Alpha	t-stat alpha	p-value	boot p-value
ABIF Norge ++	0.70	0.29	0.39	0.70
Alfred Berg Humanfond	0.31	0.21	0.42	0.82
DNB Norge (IV)	0.17	0.16	0.44	0.89
KLP AksjeNorge	0.21	0.15	0.44	0.90
Skandia Horisont	0.26	0.08	0.47	0.95
Handelsbanken Norge	0.08	0.06	0.48	0.96
Storebrand Vekst	0.06	0.02	0.49	0.96
Berg Norge gml	-0.01	-0.01	0.50	0.03
DNB Norge (I)	-0.19	-0.07	0.47	0.01
DNB Norge (III)	-0.06	-0.07	0.47	0.03
Storebrand AksjeSpar	-0.20	-0.10	0.46	0.02
Delphi Norge	-0.23	-0.12	0.45	0.04
ODIN Norge II	-0.39	-0.15	0.44	0.02
Alfre d Berg Gambak FORTE Norge	-0.38	-0.19	0.42	0.02
Nordea Kapital II	-0.71	-0.26	0.40	0.00
FORTE Norge	-1.04	-0.33	0.37	0.00
Danske Invest Norge Vekst	-0.74	-0.33	0.37	0.00
DnB Real-Vekst	-2.30	-0.36	0.36	0.00
DNB Norge Selektiv (III)	-0.38	-0.36	0.36	0.01
Nordea Norge Pluss	-0.81	-0.38	0.35	0.03
Sbanken Framgang Sammen	-1.01	-0.41	0.34	0.03
Banco Norge	-2.11	-0.50	0.31	0.00
Eika SMB	-1.23	-0.52	0.30	0.01
SEB Norge LU	-1.68	-0.53	0.30	0.01
Storebrand Norge A	-2.27	-0.56	0.29	0.01
Fondsfinans Aktiv II	-2.49	-0.65	0.26	0.00
Alfred Berg Norge Etisk	-1.23	-0.68	0.25	0.00
Arctic Norwegian Equities Class D	-1.54	-0.69	0.25	0.00
Terra Norge	-1.28	-0.69	0.24	0.01
Delphi Vekst	-1.99	-0.74	0.23	0.00
DNB Norge (Avanse I)	-1.00	-0.80	0.21	0.00
Handelsbanken Norge A10	-3.21	-0.84	0.20	0.00
NB-Aksjefond	-1.36	-0.85	0.20	0.00
Atlas Norge	-1.91	-0.91	0.18	0.00
FIRST Generator	-3.58	-0.91	0.18	0.00
Alfred Berg Aktiv	-1.71	-0.92	0.18	0.00
Alfred Berg Norge Classic	-0.94	-0.93	0.18	0.00
Nordea Barnespar	-3.02	-0.94	0.17	0.00
KLP Aksjeinvest	-2.44	-0.97	0.17	0.00
SR-Bank Norge B	-7.07	-1.05	0.15	0.00

Continued on next page

Table 12 – *Continued from previous page*

Name	Alpha	t-stat alpha	p-value	boot p-value
SR-Bank Norge A	-7.08	-1.05	0.15	0.00
Orkla Finans 30	-2.07	-1.06	0.14	0.02
RF Aksjefond	-2.32	-1.09	0.14	0.02
Postbanken Aksjevekst	-2.97	-1.17	0.12	0.01
Alfred Berg Aktiv II	-2.93	-1.20	0.12	0.02
Globus Aktiv	-6.13	-1.21	0.11	0.02
Nordea Kapital III	-3.23	-1.24	0.11	0.02
Alfred Berg N. Pensjon	-3.49	-1.24	0.11	0.03
Arctic Norwegian Equities Class A	-2.92	-1.30	0.10	0.01
Nordea Vekst	-1.77	-1.31	0.10	0.02
Storebrand Norge Institusjon	-3.56	-1.33	0.09	0.02
Arctic Norwegian Equities Class I	-2.86	-1.37	0.09	0.03
Arctic Norwegian Equities Class B	-2.99	-1.42	0.08	0.02
Plussfond	-7.01	-1.43	0.08	0.02
DNB Norge	-1.21	-1.44	0.07	0.04
DNB Norge (Avanse II)	-1.72	-1.53	0.06	0.00
Fokus Barnespar	-11.89	-1.67	0.05	0.00
Alfred Berg Vekst	-8.95	-1.72	0.04	0.00
Globus Norge II	-8.34	-1.73	0.04	0.00
GAMBAK Oppkjxf8p	-18.62	-1.81	0.04	0.00
Danske Invest Aktiv Formuesf. A	-19.06	-1.89	0.03	0.00
Globus Norge	-8.54	-1.99	0.02	0.00
Alfred Berg Aksjef Norge	-3.32	-2.03	0.02	0.01
Alfred Berg Aksjespar	-4.98	-2.14	0.01	0.02
GJENSIDIGE AksjeSpar	-4.28	-2.45	0.00	0.00
Nordea SMB	-6.44	-2.63	0.00	0.00
GJENSIDIGE Invest	-5.31	-2.96	0.00	0.00
Nordea SMB II	-16.93	-3.18	0.00	0.01
Skandia SMB Norge	-12.25	-3.22	0.00	0.10

[2] The table reports the bootstrapped results for each of the 107 funds in our sample. The funds are sorted by their alpha t-statistic. Column 2-5 reports the alpha, alpha t-statistic, parametric p-value and bootstrapped p-value. The alphas are annualized and the statistics are based on 1.000 bootstrap re-samples.

Appendix F: Cutoff t-statistics

The table present the various false discovery rate values for all the cutoff values from 1.5 to 4.0 and -4.0 to -1.5. They are presented for each assumption of outperforming or underperforming managers raging from 0 to 10. Panel A displays the result under the assumption of outperforming managers and Panel B present the result under assumption of underperforming managers.

Panel A

	-4.0	-3.9	-3.8	-3.7	-3.6	-3.5	-3.4	-3.3	-3.2	-3.1	-3.0	-2.9	-2.8	-2.7	-2.6	-2.5	-2.4	-2.3	-2.2	-2.1	-2.0	-1.9	-1.8	-1.7	-1.6	-1.5
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.03	**0.04**	0.06	0.09	0.13	0.19	0.26	0.35	0.46	0.58	0.70	0.80	0.88	0.94	0.97	0.99	0.99
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.03	**0.04**	0.08	0.13	0.20	0.30	0.43	0.57	0.71	0.82	0.91	0.96	0.98
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	**0.04**	0.07	0.13	0.22	0.34	0.50	0.66	0.79	0.89	0.95
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	**0.04**	0.09	0.17	0.30	0.46	0.63	0.78	0.89
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.03**	0.07	0.16	0.29	0.45	0.63	0.80
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.03**	0.07	0.15	0.29	0.47	0.69
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.03**	0.07	0.17	0.32	0.52
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.03**	0.08	0.20	0.37
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.04**	0.11	0.24
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.02**	0.06	0.15
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.03**	0.08

Panel B

	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	1.00	1.00	1.00	0.99	0.97	0.94	0.89	0.83	0.76	0.67	0.59	0.50	0.43	0.35	0.29	0.24	0.19	0.16	0.13	0.11	0.09	0.07	0.06	**0.05**	0.04	0.04
1	1.00	0.99	0.97	0.93	0.86	0.77	0.65	0.53	0.41	0.31	0.22	0.16	0.11	0.07	**0.05**	0.03	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
2	0.99	0.96	0.89	0.80	0.67	0.53	0.39	0.26	0.17	0.10	0.06	**0.03**	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.95	0.88	0.76	0.61	0.45	0.30	0.18	0.11	0.11	0.06	**0.03**	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.88	0.75	0.59	0.41	0.25	0.14	0.07	**0.04**	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.77	0.60	0.41	0.24	0.13	0.06	**0.03**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.64	0.43	0.25	0.13	0.06	**0.02**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.48	0.28	0.14	0.06	**0.02**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.34	0.17	0.07	**0.02**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.22	0.10	**0.03**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.13	**0.05**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Appendix G: Discussion paper - Jonas Driveklepp

Master's thesis 2021 - Discussion paper on the topic of Responsibility

The following discussion paper is based upon the master's thesis: "A performance evaluation of Norwegian mutual funds: Luck versus Skill". The main focus of our thesis is the performance of mutual fund managers in the Norwegian mutual fund market. We perform a series of tests to estimate whether we can identify any funds outperforming or underperforming their respective benchmarks. Further, we use statistical measures to determine whether the past performance of mutual fund managers originates from stock-picking skills or if it originates from luck, and in many cases, persistent luck. The analysis was performed for underperforming funds, whether it was lack of skill or simply bad luck among the managers of these funds. Thirdly, we analyze the error rates in multiple testing to create the possibility for investors to weigh their assessed cost of risk upon our analysis. Our thesis executes several statistical testing procedures upon the Norwegian mutual fund market, and some of the methods used have never before been performed onto the Norwegian market. Thus, our thesis presents novelty to the existing literature.

Responsibility in our thesis plays an important role in some of the different aspects of our thesis. I will divide the discussion into two subsections; 1) Actively managed funds 2) Statistical research.

Actively managed funds

The financial sector has experienced a severe increase in legislated responsibility towards its consumers over the last two decades. From the Financial crisis in 2008, the world's governments saw the necessity of regulations in the, up and till now, free market of immensely powerful actors (Viral V. Acharya & Matthew Richardson, 2009). The financial crisis of 2008 originated from several factors, one of them being the lack of transparency and the enormous complexity of the products offered to consumers. Mortgage bonds were one of these products (Acharya & Richardson, 2009). In essence, this was a financial product offered by banks where consumers could purchase a stake, or in many ways a bet, of a group of mortgages and their payments. The assumption being that everyone always paid their mortgages. In essence, the product was labeled as close to risk-free. The labeling of such products shows the lack of transparency towards the consumers in such financial instruments. Actively managed mutual funds are similar to these instruments. These funds proclaim that an investor's keen-eyed instincts and immense resources will have the capacity to beat the market for an extra cost of management fees. These fees are typically around 1.5% - 3.0% for actively managed Norwegian mutual funds. One of the primary responsibilities of the banks is to present these products as they are, with their inherent risk, the possibility of underperformance and the alleged certainty of outperformance of the market – to make up for the management costs. The alternative to an actively managed fund is a passively managed fund. These funds are algorithmically managed, continuously assessing the overall market and aiming to match the benchmark. The benchmark in our thesis on the Norwegian Mutual funds market is the OSEAX, Oslo All Share Index. This index reflects all stock on the Oslo Stock exchange and weights the index according to the stock's market capitalization size. Since the introduction of securities trade, there have been increased speculations; Do actively managed funds outperform their passive, less costly counterpart, and do they outperform them in such effect that the management costs incur excess return. This statement is the basis of one of our research questions. Through our analyses, and many studies before us, we come to the same conclusion; Actively managed funds do not beat their benchmark in excess of their respective management fees persistently. This result originates from several articles, Martin J. Gruber (1996), Lars Q. Sørensen (2011) and Harvey & Liu (2020), to mention a few. However, some funds do outperform in the short run, but these results are a product of luck instead of skill. Numerous articles, studies and papers have come to the same conclusion, but the amount invested in actively managed funds keeps increasing yearly. This discovery begs the question; Do the financial services providers and the banks present all available information to their customers? There seems to be an indication of a lack of transparency as scientific results differ from the actual development in the market.

Our thesis does not focus directly upon the banks' advisory strategies, nor have we assessed the specific development in the acquisition of actively managed Norwegian mutual funds. We have analyzed the historical returns in the market objectively and scientifically. We have identified evidence that an actively managed mutual fund bears a high risk for underperformance and low to non-existing possibilities of outperformance. With the increased cost of management, there seems to be little scientific evidence that consumers should choose such products. We believe this topic sheds light upon the stakeholders' responsibility and actors in the financial market, as my personal experience with investment advisory from a Norwegian bank recommends actively managed funds as a viable investment strategy. The ethical challenge in this matter is the financial actors' dilemma towards its customers. Every financial actor has stakeholders demanding growth and an increase in value. The apparent product for economic growth will be the banks' products through actively managed funds with the highest management fees. However, the best choice for the customers from our computations will be a passively managed fund. The responsibility of the bank is two-fold. They must portray ethical standards towards their customers, as well as create value for their owners. The handling of these dilemmas, I believe, can be done through transparency. As mentioned in the introduction of the paper, luck plays a big part in funds return. When buying a mutual fund, you are, in essence betting on whether that particular fund will perform in the future. The future is unknown to all. But by displaying the actively managed funds as they are, the consumer may choose a fund based on the correct assumptions. You may be lucky, and your actively managed fund outperforms – however, we can not identify any persistent skill with the managers. Or, you can invest in passively managed funds and follow the market with less risk, fewer costs, but less possibility of a luck outperformance.

Statistical research Our thesis performs several statistical procedures and combats many possible biases which may occur in multiple testing. As researchers, we have a responsibility of objectivity in our testing. A problem in general statistics is the human error and the inherent subjective aspect of testing for a result you want to achieve. General knowledge states that the first and foremost responsibility of a researcher or a scientist is to present the facts objectively and untampered. The public view of researchers is a view of credibility. If a researcher submits a result and can display full transparency and methodology, I tend to believe the work presented. Through our thesis, we have encountered this problem. Through testing, the results may often differ from your initial hopes. As a researcher, you have to accept that fact and reject the hypotheses if the numbers are insignificant. But it is not in the testing, the most prominent problems occur. The sample plays an enormous role in the statistical testing. How do you create a sample, which assumptions, when is it representative, where did the information come from, why is the sample as it is? These questions were central when assessing the sample. As a researcher, you may know what kind of sample will produce the best results for your personal ambition even before creating the sample.

The responsibility of the researcher is to remain objective. There has been notable growth in data snooping or ”p-hacking” (Chordia T., Goyal A. and Saretto A., 2017). This phenomenon is generally explained as the chase of significance or the attempt to acquire a wanted result. Many researchers are funded through an organization that wants to analyze a specific problem. Often the results are expected but must be definite for an organization to proceed with its plans or actions. In such scenarios, the researcher is not entirely objective, as his salary and the project's income rely upon an organization for funds. There may arise serious ethical challenges in such a scenario. However, the responsibility of the researcher is always to be objective. The solution to such issues is always to incorporate independent researchers to ensure unbiased results. Independency breeds objectivity. Our thesis discusses potential biases that could occur in different assumptions upon the sample and took these into account. Further, we located the sample of Norwegian mutual funds' historical returns before the research question and tests were formed to combat research bias (Sørensen, L. Q., 2011).

By writing a thesis anchored in financial parameters, we discovered that the potential ethical problems were less than a qualitative thesis might experience. Our data and analyses consist of numbers, equations and mathematical programs. We believe we experienced less responsibility for the outcome of the results and findings in the thesis due to this fact. By relying on proven mathematical and statistical approaches, and further previous research, we were able to focus on the execution of the analysis itself instead of the actual data mining. Of course, the responsibility is to hold true to the methods, numbers, and procedures, but if you are able to do this, the numbers will present the plain facts untainted. Further, one problem with statistical and mathematically proven procedures is that you must rely on the work performed by the ones before you. We have experienced through our thesis' literature review that the same tests and methods are constantly improved, contradicted, disproven or proved by other articles and researchers. This fact increases the responsibility for us to be critical to previous literature and their respective findings. To combat this problem, we implemented three well-known procedures and performed them all to present comparable results. These procedures delivered results that aligned, which in turn increase the robustness of the thesis. The responsibility of objectivity also requires the assessment of different approaches.

General discussion, conclusion and final remarks

In our master's thesis, we encountered several dilemmas of ethical nature. First, we experienced that the possibility of manipulating results would be pretty easy if the original sample were tampered with and that we could get the results we wanted. However, we discovered that the numbers and facts eventually would provide the correct picture, no matter what our priors were on the subject. Second, through the choice of sample, we encountered a lack of criticism on our part. We re-used a sample from a previous paper, which in turn was located from a reliable source. However, initially, we did not double-check the initial source of the data. As a researcher, we are responsible for the data used in our analyses. Eventually, the data background was verified and compared to the original data available from the provider to ensure its correctness.

In conclusion, our master's thesis fronts several dilemmas surrounding the term responsibility. We discover that active fund management underperforms in most cases through our analysis of active mutual fund management and its inherent results. When outperformance is present, it is the result of luck. The market characteristics do not represent these findings as there is a growth in the investment in actively managed mutual funds. This points to the question of the information flow and transparency in the actors in the financial markets and whether they perform their responsibility as objective and independent advisors to their customers, providing all available information for decision-making. Further, the responsibility of scientists is increasingly debated, as there has been an increase in data spooning and p-hacking over the years. The responsibility lies with the researchers as they are expected to perform objectively and independently in their research. However, the bias of funding in research increases the ethical dilemma of researchers. If the results align with their expected results, funding is increased; if not, funding decreases. Ultimately, our thesis projects the dilemma of responsibility through both the topic of our analysis and the methods conducted. We have experienced the necessity of responsible decision-making in our work, as we have strived to be as objective and independent we can, even though our work depends heavily on previous work and procedures. We have tried to manage these dilemmas, projecting responsibility by incorporating several approaches towards the same sample and acquired the sample and the dataset as a whole from a source independent of our thesis.

I believe we have presented responsibility towards the analyses through our actions, as previously stated. We have shed light upon several themes where a lack of responsibility may exist.

References

[Acharya, 2009] Acharya, V. V. and Richardsson, M. (2009). Causes of the financial crisis. The Critical Review, 21(2-3).

[Chordia et al., 2017] Chordia, T., Goyal, A. and Saretto, A. (2017). p-Hacking: Evidence from Two Million Trading Strategies. The SSRN Electronic Journal.

[Gruber, 1996] Gruber, M. J. (1996). Another puzzle: The growth in actively managed mutual funds. The Journal of Finance, 51(3).

[Harvey and Liu, 2020] Harvey, C. R. and Liu, Y. (2020). False (and Missed) Discoveries in Financial Economics. The Journal of Finance, 75(5).

[Sørensen, 2011] Sørensen, L. Q. (2011). Mutual Fund Performance at the Oslo Stock Exchange. The SSRN Electronic Journal.

[Ødegaard, 2011] Ødegaard, B. A. (2011). Empirics of the Oslo Stock Exchange. Basic, descriptive, results. University of Stavanger.

Appendix H: Discussion paper - André Jørgensen

Master's thesis 2021 - Discussion paper on the topic of Responsibility

This discussion paper is written in the context of the Master Thesis in a master's degree in business administration at The University of Agder. The reflection first presents a summary of the thesis, "A performance evaluation of Norwegian mutual funds: Luck versus Skill," second, we reflect on the given subject "responsibilities" in our thesis and other important aspects to the subject.

Summary of the Master Thesis

Our thesis concerns the question of whether Norwegian mutual fund managers can outperform their benchmark index. The managers need special skills to be outperforming beyond their cost of management. We also want to assess whether there is a lack of skill among Norwegian mutual fund managers. The dataset we use consists of 107 Norwegian mutual funds and their monthly returns from 1987 to 2019.

We use a pricing model with multiple factors to separate the actual performance of the managers given by an alpha (α). The results illustrate that several managers can generate positive alphas. However, we further test if this results from luck or skills because, in a large dataset, some managers will outperform or underperform by coincidence. We test this by similar approaches to the bootstrap methodology of Kosowski, Timmermann, Wermers and White (2006) and Fama & French (2010). Finally, we set the alpha to zero to operate under the null hypothesis of zero performance and calculate the p-value to check if the bootstrapped iteration generates more extreme positive values than our actual values.

The novelty of our thesis is the consideration of Type II errors and contribution by more robust conclusions than previous findings to the subject. We conduct the methodology by Harvey & Liu (2020) to test the null hypotheses that a fraction p_0 is outperforming or underperforming. In our conclusion, we cannot find outperforming managers in the Norwegian mutual fund market; however, the test power is much below the recommended level, implying that there might be outperforming managers, but our test is not powerful enough to detect them. We do, however, find statistically significant evidence that skill is absent among the worst-performing Norwegian mutual fund managers.

General Discussion

The subject of responsibility is of high importance to the field of performance evaluation within finance. In this reflection note, we evaluate the importance of responsibility towards the research process and the responsibility of the mutual fund managers. We first assess the importance of potential biases in the dataset we use. Further, we evaluate the robustness of common approaches used. We also discuss an important issue within hypothesis testing called "p-hacking." Finally, another topic relevant to our thesis is responsibility regarding mutual fund managers.

The first topic of responsibility I will discuss is the potential biases in the dataset we use. This topic concerns the probability that the dataset is not representative of the case we want to investigate. First, we will present an example of our thesis's topic if we set a minimum requirement of observations like the previous methodology by Kosowski, Timmermann, Wermers and White (2006), where the minimum is 36 observations in the dataset. This high requirement results in survivorship bias because, as we demonstrate in the thesis, the "alive" funds exhibit a higher return than the "dead" funds. Thus, we have a responsibility to include a representative dataset in the thesis, but it will not represent the actual case if we exclude all the bad performances.

The exclusion of noisy time periods also vexes empirical results. However, do this indicate that we should exclude a time period because it ruins our results? Harvey (2020) illustrated this issue in the YouTube clip "Tortured Data" with an example of the Norwegian Oil Fund. The main result was that "Abstracting from the financial crisis, we conclude that active management of both equity and fixed income has significantly contributed to the return of funds." Then Harvey raised whether they could justify excluding one of the most important financial events in modern time when evaluating financial performance. Another issue relevant to our thesis is the possibility of increasing the "test power" to accomplish the desired result. One possible variable we could change to increase the test power is by increasing the significance level that sets the Type I error. The approach of Harvey & Liu (2020) specify the importance of first setting the desired significance level to the Type I error to estimate further the Type II error (test power = 1-Type II).

There is a responsibility towards the researcher to correct the potential biases associated with multiple testing. For example, we use several tests in our thesis to determine whether there exist outperforming managers. According to Fama & French (2010), the methodology by Kosowski et al. (2006) did not account for potential correlation in the performance that the factor model could not capture. Further, they also suggested that the minimum required observation should be lower to account for potential survivorship bias in the sample. Eventually, Harvey & Liu (2020) implied that the tests by Kosowski et al. (2006) and Fama & French (2010) were not powerful to identify any possible outperforming managers. Then they suggested a methodology that tests the null hypothesis that a fraction of managers outperforms the benchmark. The availability of big data and advancing technology for computing make the importance of correct use of multiple testing a vital responsibility to use the correct testing procedure when evaluating performance.

The subject of “p-hacking” relates well to issues towards responsibilities within research approaches. “p-hacking” concerns using statistical methods to obtain a desired significant result. There are several ways to guide the research process to achieve a wanted result, and we will present the issues with some of these approaches. For example, suppose a researcher makes several statistical approaches but only reports a single “outstanding significant” result. Then, the researcher could convince that the significant result was the only result the researcher initially examined. This result would be p-hacking because the researcher does not report the other results like he initially tested to get the desired result.

Another phenomenon towards “p-hacking” was presented previously in the reflection note on the Norwegian Oil Fund, excluding important data that change the result. This exclusion of data that leads to another conclusion than if the important data was not excluded is an important part of the research process. Suppose managers want to assess their own performance and start to exclude important data to improve their own results. None of these papers would be relevant for investors to use.

There are several ways to do “p-hacking” if the researchers use any attempts to manipulate or exclude a part of the sample to achieve the most significant result. Another responsibility of concern is when researchers delegate research to an assistant. The assistant could manipulate the data to achieve a result that would please the researcher. A relationship like this could lead the assistant to p-hacking, such that the researcher would be pleased.

The primary historical responsibility for fund managers has been to generate profits for their investors. However, an increasing focus on responsibility has altered how fund managers lay their premises for the funds. There are several opportunities in the mutual fund market on funds that have clear rules towards social- and environmental responsibilities. For example, there exist funds in the Norwegian mutual fund market that only invest in climate-friendly companies or renewable energy. The demand has increased dramatically, and the fund managers have accommodated their requests by creating several opportunities for mutual funds. There are several ethical responsibilities towards investments in stocks that we will present below. First, the responsibility of children’s rights is about not investing in the companies that contribute to child labor. The responsibility towards corruption, the importance of reasonable procedures to avoid potential corruption in the companies. The responsibility for human rights is an essential aspect of investing in a company. All the mentioned strategies of responsibilities are of high importance to avoid contribution to these social problems. When assessing a potential investment, other vital aspects of responsibility are the importance of a responsible business model towards taxes and correct accounting practice.

Environmental and responsible funds have the abbreviation ESG “ Environmental, Social and

Governance." The investor needs to consider all these factors as well as the financial performance when they evaluate stocks. Nevertheless, ESG stocks do not have a clear guideline towards whether they can be named ESG stocks or not, and this makes the job harder for the investor. However, the EU wants to prepare a standard marking for these funds, EU Ecolabel. The focus on ESG stocks has increased in recent years. There seems to be a high expectation towards these stocks to do well in the future. This also leads to the problem of high valued stocks only because of their future expectations. An ESG stock could, in principle, not generate profits but still be valued extremely expensive due to their expectations.

The Norwegian Oil fund "Statens Pensjonfond Utland," which have the majority of their investment in stocks, follow strict regulations as they invest the pension of the Norwegian people. In addition, they have a council on ethics that considers the responsibility of the fund‘s investments.

Conclusion

The result of the reflection note is that both researchers and investors have an essential responsibility towards investments and their evaluation. The researcher is responsible for avoiding "p-hacking" to present reliable research results, as we illustrate by numerous examples. Further from the investor's perspective, the investor has a responsibility towards several environmental, social and governmental aspects.

References

[Fama and French, 2010] Fama, E. F. and French, K. R. (2010). Luck versus Skill in the cross-section of mutual fund returns. The Journal of Finance, 65(5).

[Harvey and Liu, 2020] Harvey, C. R. and Liu, Y. (2020). False (and Missed) Discoveries in Financial Economics. The Journal of Finance, 75(5).

[Kosowski et al., 2006] Kosowski, R., Timmermann, A., Wermers, R., and White, H. (2006). Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. The Journal of Finance, 61(6).

[Harvey and Liu, 2020] Harvey, C. R. (2020). Tortured Data.
https://www.youtube.com/watch?v=2JMqXN0pZdQ&t=1868s