Is active investing a zero-sum game?*

Markus Leippold† Roger Rueegg‡

October 27, 2018

Abstract

To study the hypothesis whether active investing is a zero-sum game, we analyze the alpha of active and index mutual funds from a global sample of more than 60,000 equity and fixed income funds. Using a new robust statistical test, we cannot reject this hypothesis for the vast majority of investment categories. We also find that the average active fund has less exposure to traditional risk factors, but higher sensitivity to alternative risk premia. Fund persistence and the impact of size and fees adds further support to the hypothesis.

JEL classification: C12, G10, G11, G20, G23

Keywords: Active investing, index investing, mutual funds, robust alpha test

*We are indebted to Fabian Ackermann, Andreas Blöchlinger, Vlad Dubikovskyy, Michel Habib, Thorsten Hens, Emmanuel Jurczenko, Andreas Kappler, Robert Kosowski, Harald Lohre, Per Ostberg, Julien Penasse, James Sefton, and Michael Wolf for helpful comments. We thank the participants at the Frontiers of Factor Investing Conference 2018 at the University of Leicester, and the seminar participants at the University of Zurich, Zurich Cantonalbank, Unigestion Geneva, at the 2018 UBS Quant Conference in London, the 2018 Financial Data Science & Econometrics conference at the University of Loughborough, and the Paris December 2018 Finance Meeting EUROFIDAI - ESSEC. The content of this paper reflects the personal view of the authors. In particular, it does not necessarily represent the opinion of Zürcher Kantonalbank. This paper was previously circulated under the title “Fifty Shades of Active and Index Alpha.”

†Department of Banking and Finance, University of Zurich, Plattenstrasse 14, 8032 Zurich, Switzerland, email: markus.leippold@bf.uzh.ch.

‡Department of Banking and Finance, University of Zurich and Zurich Cantonal Bank, Hardstrasse 201, 8005 Zurich, Switzerland, email: roger.rueegg@zkb.ch.

1 Introduction

The emergence of index investing has led to a seemingly endless debate about the merits of active portfolio management. Many research papers, investors, and advisors place themselves in either the active or passive camp. The staunch defenders of active investing argue along the lines of Berk and Green (2004), who show that rational markets do not contradict the existence of skilled fund managers who consistently beat the market. They build their argument on a basic principle of economics: agents earn economic rents if, and only if, they have a competitive advantage. Hence, active investing is a zero-sum game after fees. Recently, Berk and van Binsbergen (2015) have provided empirical support for the claim that mutual fund managers do have skills.

In contrast, the proponents of passive investing argue along the lines of Fama and French (2010) in that the high fees of active management turn it into a negative-sum game after costs. Indeed, French (2008) and Fama and French (2010), among many others, provide ample evidence that actively managed US equity mutual funds underperform their multi-factor benchmark after fees. In their view, active investing is at most a zero-sum game before fees, but definitely not after fees. Consequently, over the recent years, we have witnessed a massive inflow of funds into index investing. These observations naturally drive us to question the value of active management.

According to the logic of Sharpe (1991)'s active management arithmetic, active investing is doomed in aggregate, as French (2008) puts it. However, to escape this seemingly irrefutable conclusion, Sharpe (1991) leaves a back door open by pointing out three potential flaws in his theory. First, passive managers might not be truly passive. Second, there might be substantial differences among active managers.[1] Third, the summary statistics of active managers might not truly represent the performance of the actively managed dollar. In our analysis, we shed light on these three potential pitfalls by first applying a factor analysis not only on active but also on index fund. Second, we account for the heterogeneity of active asset managers by differentiating between institutional and

[1] For example, Garleanu and Pedersen (2018) consider a model where managers with larger and more sophisticated investors are expected to outperform. Their theoretical model is supported, e.g., by the findings of Gerakos et al. (2016), who show that institutional investors outperform their strategy benchmarks after fees. Their analysis is based on self-reported but GIPS (Global Investment Performance Standard) compliant data, which still may inherit some biases, while our analysis is based on publicly available performance data. In addition, Pastor et al. (2015) argue that the higher competitions in big active mutual fund industries decrease the fund's ability to outperform passive benchmarks.

retail funds, equity and fixed income funds, geographical regions, and investment categories. Lastly, we analyze the value-weighted performance before and after fees, and we benchmark against multi-factor models and investable indexes. For this extensive study, we include 61,269 equity and fixed income funds that held USD 17.8 trillion assets under management by the end of 2016, thereby substantially increasing the power of our tests.[2]

Since our preliminary data analysis indicates that there are both serial and cross-sectional dependencies in our fund data, we have developed a robust test for the manager's alpha, defined as the excess return relative to an appropriate benchmark. Our test is robust in the sense that it takes into account potential serial dependence in mutual fund returns.[3] We can then use these robust test statistics for the alphas as input for the appropriate multiple hypothesis adjustments.

As Berk and van Binsbergen (2017) convincingly argue, there is no unique way to measure performance: it depends on the research question. If we want to assess the rationality of fund investors and the degree of competition in different markets, then the appropriate measure is the fund's net alpha. If, however, we want to measure the manager's skill, then the appropriate measure is the value-weighted gross alpha, or if we want to test whether active and index investing is a zero-sum game after costs, then we must use the value-weighted net alpha. Unlike most of the previous studies, which prominently focus on active US mutual funds without differentiating between retail and institutional funds, we use the richness of our dataset to explore the performance of active funds from many different perspectives.[4]

The choice of benchmark is just as critical as the measurement of performance. Typically, researchers use a well established multi-factor model to proxy for the alternative opportunity set available to investors. However, multi-factor models include long–short portfolios with often very high turnover, generating considerable transaction costs. Furthermore, also as argued by Berk and van

[2]Of these funds, 56,136 are actively managed, and 5,133 are index funds.

[3]In a simulation study, we find that the conventional inference techniques are liberal in rejecting the null hypothesis, while we observe still liberal but accurate empirical rejection probabilities for our robust alpha test based on block resampling. Other papers that conduct bootstrapped inference, such as Kosowski et al. (2006) and Fama and French (2010), sample one-period returns and, therefore, lose any information about the potential dependence over time. Hence, the simulation results convince us that our statistical framework is the appropriate choice to carry out our research task.

[4]Recently, Ferreira et al. (2013) analyze the performance of 16,316 open-end actively managed equity funds in 27 countries from 1997 to 2007. They find that equity mutual funds around the globe, in general, underperform. Banegas et al. (2013) focus on 4,200 European equity mutual funds and find that European equity funds outperform the market.

Binsbergen (2017), we might have a situation in which we are measuring the performance of a fund at a time when the fund manager would not have known about some factors, as they were identified only much later.

Nevertheless, Fama and French (2010) argue that benchmarking against multi-factor models leads to the same conclusions as benchmarking against index funds, because the value-weighted portfolio of index funds exhibits close to zero alphas. Although we agree with their arguments for the US equity market, our single fund analysis reveals that many index funds also exhibit negative alphas relative to a multi-factor benchmark, depending on the asset class and the market. Also, there is no general agreement on the factors that should be included in a benchmark, leading to a severe selection bias. Therefore, to measure fund performance, we abandon multi-factor models as a benchmark in favor of suitably defined investable benchmarks, which allow a fair comparison of active and index funds. Of course, we still need multi-factor models to understand the potentially different risk and style exposures of active and index funds.

Under the assumption that index investors try to replicate the market portfolio and believe in the efficient market hypothesis, we compare the average dollar weighted return of active investors that build an opinion with the average dollar weighted return of an investor who has no opinion about the securities within a certain investment category.[5] Using such an approach, we also take into account the costs of replicating the market portfolio that arise due to management fees, transaction costs, and we also guarantee that factors with a low market capitalization receive a lower weight.

In our preliminary analysis, we find significant evidence for serial and cross-sectional dependence in our mutual fund data. Therefore, we use a statistical framework with two key elements. First, we develop a robust statistical test for the mutual funds' alpha, which takes into account serial dependence. Second, we use these test statistics as input for the multiple hypothesis testing methods of Barras et al. (2010) and Romano and Wolf (2016), which are robust to the presence of (mild) cross-sectional dependence. In a simulation study, we find that the standard inference techniques are liberal in rejecting the null hypothesis, while we find accurate empirical rejection probabilities for

[5]Similarly, Fama and French (2010) use such an equilibrium accounting perspective to argue that the actively managed mutual fund industry does not cover the costs they impose on investors. However, they only concentrate on one portfolio, formed by value-weighting the funds in the "Active US Equity" category, and compare it to the market portfolio.

our block resampling based alpha test. Other papers that conduct bootstrapped inference, such as Kosowski et al. (2006) and Fama and French (2010), sample one-period returns and, therefore, lose any information on the potential serial dependence; nor do they allow for multiple hypotheses.[6]

When applying our robust multiple hypothesis alpha test to single funds against multi-factor benchmarks, we find that a large fraction of active equity mutual funds deliver negative alphas after fees. Therefore, we provide international evidence for the results of Fama and French (2010).[7] Surprisingly, however, when we apply the same tests to index funds, we find that they also show negative alphas after costs. While most of the literature concentrates on the equity market, we also conduct a multi-factor analysis of the fixed income mutual funds. In contrast to the equity market, we find that there is a substantial portion of active USD fixed income funds with a positive alpha.

Given that we observe substantially negative alphas for index funds in both the equity and fixed income markets, we question the plausibility of using multi-factor benchmarks.[8] Therefore, in a further step, we analyze the net alphas of single active funds against investable benchmarks. We find that most of the active funds exhibit zero alpha. While for institutional equity funds in the US, we find a negligible proportion of negative-alpha funds, this proportion is higher for retail equity funds. Again, we observe a large fraction of index funds with negative alphas under an investable benchmark. We suspect that this negative performance may be caused by the negative performance of small funds.

Since Berk and Green (2004) argue that fund size is a crucial element in the analysis of fund performance, we value-weight the alpha of active funds within the Morningstar investment categories. The multiple hypothesis test of Barras et al. (2010) cannot be applied to value-weighted alphas. Therefore, we switch to the method of Romano and Wolf (2016). We find that there are significant negative alphas after cost for the institutional "US Equity Large Cap Blend" and retail "Canada Fixed Income" categories. This finding corroborates the conclusion of Fama and French (2010) for these specific investment categories. However, for all of the other categories, our results support Berk and

[6]The Matlab code of the robust statistical framework for the alpha is available from the authors on request.

[7]In particular, we find the highest proportion of funds with positive alphas against the multi-factor benchmark for institutional investors to be in Europe and Japan, which confirms the findings of Banegas et al. (2013).

[8]Index funds should have zero alpha on average if the multi-factor benchmark were appropriate. Of course, there is still an on-going debate about which multi-factor model best describes the investment opportunity set. For our analysis, we relied on the most common models.

Green (2004) and Berk and van Binsbergen (2015) in that we cannot reject the hypothesis that it is a zero-sum game after costs.[9] Furthermore, we see periods in which the average active managers underperform the index alternatives, such as, before the dot-com bubble burst, during the financial crisis, or in the very recent period from 2014 to 2016. However, we can also observe periods in which active managers, on average, outperform, such as, from 2000 to 2007 or from 2009 to 2014.

Analyzing the drivers of the difference in the performance of active and index funds, we find that the equity and fixed income active managers have less exposure to traditional risk factors such as market and duration risk. Instead, active equity funds have a small cap and growth stock bias and active fixed income funds load on credit risk. Surprisingly, when the market is affected by unexpected volatility shocks, active management tends to underperform the average index investor. In periods of calm markets and when the implied volatility decreases, active managers tend to outperform. We explain this finding as being due to active managers who prefer to sell insurance and generate exposures to risk premia that perform well in good times but may cause substantial losses in bad times. The significant higher exposure to small cap companies for equity and credit risk for fixed income managers supports this hypothesis.[10]

Our data also allows us to shed light on the difference between retail and institutional funds. We find that after fees, there are a majority of unskilled mutual funds for the retail segment. In contrast, we see a more balanced proportion of skilled and unskilled funds for the institutional sectors and outside of the US. Thus, our results provide direct evidence for Garleanu and Pedersen (2018), who argue that more sophisticated investors outperform small investors because of the higher economies of scale in searching for skilled active managers.[11] Moreover, our results endorse the hypothesis of Gennaioli et al. (2015), who claim that trust is an essential component of the high fees in asset

[9] Furthermore, our findings resonate well with Pastor et al. (2015), who argue that in markets in which the mutual fund industry is big, such as the "US Equity Large Cap Blend," active alphas tend to be negative, and the equal-weighted alpha within investment categories exceeds the value-weighted alpha.

[10] Agarwal and Naik (2004) find similar return patterns for hedge funds. Thus, mutual funds try to profit from the same opportunities as hedge funds but have of course a narrower set of investment opportunities, due to regulatory restrictions. However, our results seem to contradict some of the previous findings, such as those of Moskowitz (2000) and Kosowski (2011), among others. They find that actively managed mutual funds tend to perform better than their passive benchmarks in bad times. However, these papers do not cover the recent financial crisis.

[11] Also, our empirical analysis supports the findings of Gerakos et al. (2016), who show that institutional investors outperform their strategy benchmarks after fees. Their analysis is based on self-reported but GIPS (Global Investment Performance Standard) compliant data, which still may inherit some biases, while our analysis is based on publicly available performance data.

management, and who argue that active retail managers profit from pandering to trusting investors by buying hot assets, which explains the tendency for active retail mutual funds to have positive exposure to growth stocks.

We further demonstrate, along the lines of Carhart (1997), that the average active retail investor can significantly improve their performance over the period ranging from 1993 to 2016, provided the worst-performing active mutual funds of the past year are neglected. However, when the investor concentrates only on the top performing funds, the overall performance cannot be significantly improved. In addition, we explore the role of fund size and fees on fund performance. Sorting active fund portfolios according to their performance persistence, fees, and size, we find that winner portfolios with low-fee and small funds tend to outperform but their alpha does not survive our test statistics. However, for both equity and fixed income retail funds, we find that a fund investor is well advised to avoid high-fee and small losers, as they generate significantly negative alphas.

The remainder of this paper is organized as follows. Section 2 discusses the data and performs a preliminary analysis, which motivates the design of our empirical tests. Section 3 presents our robust alpha test and the multiple hypothesis framework. In Section 4, we first compare the performance of single index and active mutual funds when benchmarked against factor models and an investable index. In Section 5, we provide a comparison of the value-weighted performance of active and index mutual funds portfolios across investment categories and asset classes. We analyze the drivers of the difference in the performance of active and index funds, and we explore the role of performance persistence, fund size, and fees. Section 6 concludes.[12]

2 Preliminary analysis

After describing our data, we analyze the potential time and cross-sectional dependencies in mutual fund returns to guide the formulation of appropriate test statistics for our hypotheses.

[12] All Matlab code used in this paper is available from the authors on request.

2.1 Data

Our mutual fund sample is drawn from the Morningstar database and ranges from December 1991 to December 2016. We include a total of 61,269 funds from different asset classes.[13] Table 1 shows the summary statistics of cross-sectional monthly attributes across asset classes. For the active funds, we analyze in all 14,969 institutional and 46,300 retail funds, while we have 56,136 active and 5,133 index funds. In general, there are fewer index funds, but they show higher average total net assets (TNA) and net returns, and also lower fees and about the same average years in the database. As expected, the institutional funds charge lower fees than their retail counterparts.

[Table 1 about here.]

As of December 2016, the total net assets of equity retail funds amounted to USD 9 trillion, those of fixed income retail funds to USD 3.7 trillion, and those of equity institutional funds and fixed income institutional fund to USD 3.1 trillion and USD 2 trillion, respectively. Since institutional investors often invest their money through mandates, there are fewer institutional funds than retail funds. The assets under management for index funds have been steadily increasing over our sample period. By the end of 2016, we find the highest concentration of index funds for equity funds, with 28% for retail and 32% for institutional funds. Looking at the fixed income funds, we find 18% of the retail and 13% of the institutional funds were index funds. For a more detailed description of the data and the data cleaning procedures, we refer to Appendix A.

2.2 Dependency analysis

It is well known that statistical inference for econometric models is severely complicated by the existence of serial and cross-sectional dependencies. Fama and French (2010) find that cross-sectional dependence can materially change the inference and, therefore, propose an appropriate adjustment for their single fund analysis.[14] At the same time, they correctly point to a potential caveat in

[13] In comparison, Pastor et al. (2015) explore 3,126 actively managed US equity-only mutual funds while Berk and van Binsbergen (2015) use 5,974 actively managed funds. Hence, we add to the existing literature by providing evidence based on our new dataset. Furthermore, to the best of our knowledge, we are the first to apply a robust multiple hypothesis framework to active and index mutual funds in an international context.

[14] See also Chen et al. (2017).

their resampling approach. Because they perform a random sampling of months, they lose any effects of autocorrelation. Similarly, neither does Barras et al. (2010) take into account serial dependence, claiming that they find such an effect only for a few mutual funds.[15] Moreover, while high cross-sectional dependencies could potentially bias their estimators, they find a low average pairwise correlation of 0.08 in their sample and argue that the cross-sectional dependencies are sufficiently low to allow consistent estimators. Given that we analyze not only the returns of single funds as in Barras et al. (2010) and Fama and French (2010) but also of mutual fund portfolios, and compare them against multi-factor and investable benchmark models, we must test for time dependence in a variety of settings. Therefore, we take a closer look at our data.

To compare active and index investing, we construct two types of different benchmark models. First, we apply the commonly used multi-factor models. In particular, for the equity analysis, we use the regional five-factor model with "market," "size," and "value" factors as given in Fama and French (1992) and add the "momentum" factor of Jegadeesh and Titman (1993) as well as the "betting against beta" factor of Frazzini and Pedersen (2014).[16] For the fixed income analysis, we apply a four-factor regional model including the "shift," "twist" and "butterfly" factors, as well as MSCI Inc.'s credit risk factor, measured as the BBB–AAA spread.[17] Second, we focus on the investable one-factor benchmark model, which we build as the value-weighted return of the index funds within the investment category of the analyzed time-series.[18]

We first test for serial dependence, applying the classical Ljung–Box (LJ) test and, as a robustness check, the distribution-free test of Genest and Rémillard (2004).[19] For both tests we must fix the number of lags L, for which we use the automatic block-length selection for the dependent bootstrap of Politis and White (2004) and the correction of Patton et al. (2009). We find that most mutual funds show an optimal block size of two or three. Therefore, we set the lag L to three for the two

[15] However, recently, Zhang and Yan (2018) find that the standard bootstrap can be misleading.

[16] The regional factors were retrieved from the homepage of Kenneth French, while the "betting against beta" factor is provided for each region on the homepage of AQR.

[17] The factors "shift," "twist," and "butterfly" represent the risk for a change in the level, steepness, and curvature of the term structure. See DeMond et al. (2012) for a description of the factors.

[18] For the investable one-factor model, we require at least 12 monthly returns and for the multi-factor models at least 36 monthly returns.

[19] Much criticism has been leveled at the possible low power of the LJ test. The LJ test is based on autocorrelations and, hence, it is not a real test of independence. The test developed by Genest and Rémillard (2004) uses ranks and, therefore, is distribution-free and does not depend on the underlying distribution of the observations.

tests.

[Table 2 about here.]

Table 2, Panel A, presents the number and percentage of funds that have a p-value below 5% based on the standard LJ test and the test of Genest and Rémillard (2004). Both tests provide us with a similar pattern. We find that for single equity funds, the percentage of rejected null hypotheses of no serial dependence over time ranges from 14% to 23%, with a slightly higher rejection rate for retail funds. Similarly, the percentage of fixed income funds rejecting the null ranges between 12% and 22%. For fund portfolios, the rejection rates can even be as large as 80%, although the number of portfolios is rather small and, hence, the results have to be interpreted with care. Nevertheless, it becomes clear that single fund and portfolio residuals are not serially independent. Furthermore, when we go into more detail, we find that the single mutual funds with the longest available time-series show a higher percentage of rejections. For example, the 2% oldest equity and fixed income single mutual funds, benchmarked against the investable model, exhibit statistically significant serial dependence in 40% and 63% of the cases.[20] Overall, signs of serial dependence can be found in roughly every fifth single mutual fund, and every third mutual fund portfolio. This evidence clearly justifies the need to control for dependence over time when we analyze the alpha of single and portfolios of mutual funds against different benchmark models.

We next test for cross-sectional dependence, which might occur if mutual funds "herd" in their holdings, as is shown by Wermers (1999). To detect cross-sectional dependence in our data, we apply the test of Pesaran (2004).[21] To compute the test statistic, we concentrate on funds that have more than one time period in common. Panel B of Table 2 presents the average pairwise correlation of the residuals together with the p-values of the Pesaran (2004) test. We find that we can reject the null hypothesis of no cross-sectional dependence at the one percent significance level for all single and portfolio fund categories.

When we compute the average pairwise correlation in our single fund sample, we find the same

[20] We do not report these more detailed results here, but they can be obtained from the authors.

[21] Compared to the well-known Lagrange multiplier test of Breusch and Pagan (1980), this test is correctly centered for a large sample and comparably short time-series, which is precisely the case here, since we have a broad cross-section of mutual funds but a comparably small time-series. Also, Pesaran (2004) finds that the test has satisfactory power even under weak cross-sectional dependence.

value for our US equity fund sample as Barras et al. (2010), i.e., an average of 0.08. For the single mutual funds with the investable benchmark models, we find average correlations between 0.04 for retail equity and 0.15 for institutional fixed income funds. For the multi-factor benchmarks, the correlation turns out to be considerably higher: 0.22 for retail equity funds and 0.62 for retail fixed income funds. On the portfolio level, we find average pairwise correlations of 0.13, 0.09, and 0.43 for the investment categories, equity, and the fixed income mutual fund portfolio, respectively. Although Barras et al. (2010) are not overly concerned with cross-sectional dependencies, we cannot merely use their reasoning given the elevated levels of average pairwise correlation, especially for fixed income funds.

3 Robust alpha test and multiple hypotheses

The above empirical evidence dictates that statistical tests must take into account both serial and cross-sectional dependence. To control for serial dependence, we propose a robust alpha test based on a studentized block bootstrap, which improves the accuracy of an inference for dependent time-series data compared to other methods.[22] To compute the bootstrapped t-statistics and p-values we closely follow Ledoit and Wolf (2008, 2011), who study the related problem of testing whether two Sharpe ratios or two variances are equal. We outline the mathematical details of the bootstrapped standard error of the estimated alpha in Appendix B. Once we have calculated the bootstrapped t-statistics and p-values in Equations (B.12) and (B.13), we can use them as input for multiple hypothesis testing.

While we control for serial dependence for the single-hypothesis alpha test, we control for cross-sectional dependence for the multiple-hypothesis method. Depending on whether we analyze single funds or portfolios of funds, we control either the false discovery rate (FDR) or the family-wise error rate (FWER).[23] As Bajgrowicz and Scaillet (2012) argue, investors do not rely on a single active manager but instead diversify between different managers. Therefore, in their view, it is favorable to control the amount of falsely rejected hypotheses (FDR) instead of investing only in the best

[22] See, e.g., Lahiri (2003), Haerdle et al. (2003), and Ledoit and Wolf (2008, 2011).

[23] The FWER dates back to Bonferroni (1936), and is defined as the probability of at least one false discovery. Romano and Wolf (2005a,b) introduce a stepwise multiple testing procedure that not only has higher statistical power than the tests of Bonferroni (1936) and Holm (1979) but also allows for cross-sectional dependence. The FDR is defined as the expectation of the proportion of falsely rejected null hypotheses. For a larger number of hypotheses, Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) show that it is favorable to control for the FDR.

strategies, as is the case when controlling the more conservative FWER. This diversification argument of Bajgrowicz and Scaillet (2012) no longer holds for portfolios of funds. Therefore, we control for the FDR when analyzing single funds and for the FWER when analyzing portfolios of funds.

For the FDR, we rely on the approach of Barras et al. (2010). A strength of their approach is that it regards each fund in isolation. However, this advantage comes at the cost that a high cross-sectional dependence could potentially bias their estimators. Thus, we conduct a set of Monte Carlo simulations for the multi-factor model in the fixed income market, where the findings in Panel B of Table 2 have indicated a high degree of cross-sectional dependence. In unreported results, we find that the average estimates are less stable but still close to the estimates where we assume independent residuals.[24] Therefore, we are confident that, for our application, we have consistent estimators also for the markets with a higher degree of herding. However, we must be aware that for the single fixed income market outside of the US, we have a higher estimation error.

While for the analysis of single funds, we have a large number of hypotheses, we only have a few hypotheses for the analysis of portfolios of mutual funds. Hence, we prefer to control for the FWER, applying the state of the art multiple hypothesis framework of Romano and Wolf (2016), which provides an efficient way to calculate the adjusted p-values. Since for the fund portfolios we have no missing values or disconnected time-series, as is the case for single funds, we can jointly sample blocks of fund and benchmark returns, thereby taking fully into account cross-sectional dependence. As for the FDR, we sample the test statistics with our robust alpha test, which allows us to take into account the serial dependence structure.

[Table 3 about here.]

Table 3 summarizes the motivation for our estimation strategy. Our test based on the block bootstrapped alpha is, in combination with the FDR (for single funds) and the FWER (for fund portfolios), a suitable method taking into account both serial and cross-sectional dependence simultaneously, as evidenced by our preliminary analysis. Also, both frameworks take into account the characteristics of the data. For the single fund analysis with a large cross-section and a small overlap of the time-series, we regard each fund in isolation and therefore prefer to control the FDR. For the

[24]These results can be obtained from the authors.

portfolio analysis with a small cross-section and fully connected time-series, we focus on the more restrictive FWER and jointly block bootstrap the entire data sample. To explore the accuracy of our test, we present the results of a simulation exercise in Appendix C. We find that our robust alpha test is still liberal but more accurate since it also corrects for the serial dependence observed in the data. The standard inference tests are too liberal in rejecting the null hypothesis. Thus, when we apply the standard tests or sample only one return each time instead of a block of returns, we generate more type I errors (false positive findings) than expected by the test.[25]

4 Single mutual funds

We first analyze the distribution of the single fund alphas measured against the regional multi-factor benchmark models. While such a comparison is not suited to identifying skilled managers, it gives us an idea about the risk drivers and style exposures of the different funds. Later, we compare the single funds' alphas benchmarked against an investable index.

4.1 Multi-factor benchmark

For equity funds, the five-factor model is used as the multi-factor benchmark. It is based on the three-factor model of Fama and French (1992) but includes the "momentum" and "betting against beta" factors. For fixed income funds, we rely on MSCI's four-factor model with the corresponding regional factor returns. As the first step, we calculate our robust alpha, where for each individual fund, we apply the optimal block size with the method of Politis and White (2004) and Patton et al. (2009).[26] For the multiple-hypothesis adjustment, we control for the FDR using Barras et al. (2010) and compute the proportion of negative, zero, and positive alphas after fees. It is important to emphasize that we equal-weight each fund, since the method of Barras et al. (2010) regards each fund in isolation and does not allow value-weighted adjustments. Table 4 reports the results.

[Table 4 about here.]

[25] We also note that even if there is no serial dependence, our block-bootstrapped alpha test statistic is accurate.

[26] For robustness, we also applied a block size of six, which yields the same results.

For equity funds, we find that the proportion of active funds with zero alpha is 62.3% for retail and 77.4% for institutional funds. As expected, the proportion of zero alpha funds for index investors is higher, at 68.4% for retail and 79.5% for institutional funds. We also find that the percentage of funds with a significantly positive alpha is the highest for active institutional funds, 3.5%, while for all the other categories there are between 1.4% and 1.9% single mutual funds with a positive alpha. The proportion of funds with a negative alpha is the highest for active retail funds, 35.9%, while active institutional funds have 19.1% of their managers generating a negative alpha. Surprisingly, we also find that 29.7% of the retail and 19.1% of the institutional index funds provide a negative alpha. At the same time, we observe 19.1% of the institutional active funds with a negative alpha. Hence, for institutional funds, we put both active and index managers at a similar disadvantage by using a multi-factor benchmark if we were to interpret the resulting alpha as skill. Moreover, for the US market, there are more institutional index funds (33.1%) than institutional active funds (30.7%) having a negative alpha.

Focusing on US institutional active equity funds, we find that our 69.3% zero funds compares well with the 75.4% of Barras et al. (2010, Table II). At the same time, we have a higher fraction of negative alpha funds, 30.7% against their 24%. Considering the impact of luck in the left tail, however, the proportion of significantly negative alphas (*FDR 10 alpha* < 0) drops to 3%, which is considerably smaller than their 13.6%. Interestingly, the proportion of significantly negative alphas for index funds only drops to 17.2%. From this perspective, index funds seem to perform even worse than active funds, when benchmarked against a multi-factor model.

For fixed income funds, we observe that for retail active funds, there is an equal number of negative and positive alphas (around 16%), while for retail index funds we find a smaller fraction of negative alpha funds in favor of a higher fraction of zero alpha funds. For institutional funds, we find more positive than negative alpha funds. Hence, as with equity funds, retail funds seem to perform worse than institutional funds if benchmarked against multi-factor models. For the regions outside the US, we do not find active funds in the fixed income universe with a negative alpha. For the US, the shares of negative, zero, and positive alphas are of similar magnitudes.

While these comparisons are informative about the style and risk exposure of the different funds, they neither represent manager skill since the alphas are not value-weighted, nor do they provide

useful information for return-chasing fund investors since multi-factor models provide only an unfair benchmark.[27] The inappropriateness of multi-factor benchmarks for performance measurement becomes most evident from the observed negative alphas of the index funds. If multi-factor benchmarks were fair, then index funds should have zero alpha on average.

4.2 Investable benchmark

Given the above concerns, we next construct investable benchmarks based on Morningstar's investment categories. We rely on these categories as they are well established in the industry, and their definition perfectly serves our intention to benchmark active funds.[28] To construct an investable benchmark, we value-weight all index funds within a given category. By value-weighting the index funds in each category, we obtain the investable benchmarks which we use for calculating the alphas of active funds. To make the analysis comparable to the previous section, we only include those Morningstar's investment categories that include the investment regions US, Global, Europe, Japan, and Asia ex-Japan for the equity funds. For the fixed income market, we include the categories in US dollar US, Swiss Franc CHF, Europe EU, and Sterling GBP. We first calculate the p-values from our robust alpha test, after fees and with active funds benchmarked against the corresponding value-weighted category index. We then compute the estimated percentage of negative, zero, and positive alpha funds using the method developed by Barras et al. (2010).

[Table 5 about here.]

Table 5 shows the results using the same categorization as in Table 4. Strikingly, we find much higher averages of zero alpha active funds for both retail and institutional funds. In particular, for the US and the Global categories, the difference is substantial. For instance, with an investable index as benchmark, the fraction of zero alpha active US institutional funds rises from 69.3% to 80.1% and the proportion of significant negative alphas decreases from 25.3% to 0.2%. At the same time, the

[27] See, Berk and van Binsbergen (2015).

[28] We acknowledge that there are many routes to take for benchmarking fund portfolios. In practice, when investors or active managers focus on a specific investment category, they do not compare themselves with the multi-factor models in general, rather, they compare themselves with other funds within the same category. As Morningstar states on its website, "the classifications were introduced in 1996 to help investors make meaningful comparisons between mutual funds." While the investment objective stated in a fund's prospectus does not always reflect how the fund actually invests, Morningstar places funds in a given category based on their portfolio statistics and securities holdings.

proportion of significantly positive alpha funds remains at 0.0%. Hence, the large fractions of zero alpha funds after fees support the equilibrium argument of Berk and Green (2004).[29]

What surprises us in Table 5 is the remarkably large fraction of index funds with a negative alpha, especially in the US. The average of zero alpha index funds is below the average of zero alpha active funds. This observation may be due to the fact that, so far, we have ignored fund size in our analysis, since the framework of Barras et al. (2010) does not allow us to value-weight the funds's alphas. Fund size, however, is a crucial element in the argumentation of Berk and Green (2004). Therefore, we now analyze the funds' performance while taking into account fund size.

5 Value-weighted portfolios of mutual funds

As Berk and Green (2004) argue, funds managed by skilled managers attract greater portfolio flows than funds managed by unskilled managers. Hence, if we want to measure the skill of a fund manager, or if we want to test whether active investing is a zero-sum game, we must measure performance on a value-weighted basis and against an investable benchmark.[30]

5.1 Investable benchmark

For the performance analysis of fund portfolios, we focus again on the same investable benchmarks as we used in Section 4.2. Since we require connected time-series for our multiple hypothesis adjustment, we focus on the periods from 1993 to 2016 and 2000 to 2016, which allows us to include more investment categories for the more recent time periods. Given that index mutual funds only emerged recently, we observe for the period starting in 1993 at least one index fund for four institutional and 17 retail categories. For the more recent period starting in 2000, we obtain 30 investment categories

[29] Also, in Table 4, the proportion of active fixed income funds with significantly positive alphas is surprisingly high. With the investable index as benchmark, these numbers turn out to be much more moderate, pointing at the potential problem of defining appropriate multi-factor benchmarks. Hence, switching to an investable benchmark allows a much more realistic assessment of actively managed funds.

[30] As an additional exercise, we also benchmarked our portfolios of funds against multi-factor models using the same categorization as in the previous section. For active institutional US equity funds, our results show a significantly negative alpha and, therefore, are in line with Fama and French (2010). However, except for active institutional Global funds, all other active alphas are insignificant. In contrast, we find again some significantly negative alphas for index funds. For fixed income funds, we find for active institutional funds a significantly positive alpha. Since we have argued above that multi-factor benchmarks are not suitable for performance measurement, we do not report these results here.

for the retail segment and 12 investment categories for the institutional segment. Hence, we end up with 63 categories. By value-weighting the index funds in each category, we obtain the investable benchmarks which we use for calculating the alphas of active funds.

[Figure 1 about here.]

In Figure 1, we plot the robust p-values against the net and gross alphas for each of the available investment categories. As argued in Section 3, we adjust the p-values for multiple hypothesis testing using the method of Romano and Wolf (2016). After fees, we find the "US Equity Large Cap Blend" category for institutional funds and the "Canada Fixed Income" category for retail funds to significantly underperform the alternative of the value-weighted index funds for both periods. For the negative alpha of the "Euro Fixed Income" retail category and the period from 1992 to 2016 we also find a significant p-value. Hence, only for three investment categories can we reject the zero-sum game hypothesis of Berk and Green (2004). Furthermore, our finding that "US Equity Large Cap Blend" institutional funds underperform after fees is perfectly in line with the argument of Pastor et al. (2015), in that higher competition in big active mutual fund industries leads to diminishing returns to scale. Before fees, there are no investment categories with significantly negative alphas. However, we find "US Fixed Income" from 1992 to 2016, and "Global Equity Large Cap," "Emerging Markets Equity," and "Europe Equity Large Cap" from 2000 to 2016 for institutional, and also "Global Equity Large Cap" from 1992 to 2016 and "Global Equity" from 2000 to 2016 for retail funds to significantly outperform the value-weighted index funds.[31]

In Figure 2, we show the cumulated aggregated alpha of active index funds over time.[32] Equal weighted, all investment categories for equity funds provide a positive alpha over time, even after fees. Value weighted, the aggregated alphas remain positive before fees, but they are zero for institutional funds and slightly negative for retail funds after fees. For the fixed income mutual funds, we find that the value-weighted alpha across investment regions is positive for the equal and value-weighted

[31] We remark that the choice of a block size of three is a conservative choice. As a robustness check, when we apply a block size of six or nine, the p-values increase slightly. The "Euro Fixed Income," "Emerging Markets Equity," and "Europe Equity Large Cap" categories, which all exhibit a p-value just below 10% for the block size of three, start to show insignificant p-values, further supporting the theory of a zero-sum game after fees.

[32] We first compute the value-weighted alpha within an investment category against the value-weighted benchmark of the index funds and then aggregate the investment categories with equal and value-weights. We also split into retail and institutional funds before and after fees. For the equity mutual funds, we find that active mutual funds provide a superior alpha than index funds in every analysis before fees.

aggregation of the investment categories. For the institutional funds after fees, we also observe positive alphas over time. There are three major periods where active managers underperformed their index counterpart: equity funds before the burst of the dot-com bubble, both equity and fixed income funds in the financial crisis, and a slight underperformance in the recent past, especially after fees.

[Figure 2 about here.]

The fact that the equal-weighted alpha across investment regions is higher than the value-weighted alpha adds evidence to the theory of Pastor et al. (2015) in the sense that the higher competition in big active mutual fund industries decrease the fund's ability to outperform passive benchmarks. For the fixed income mutual funds, however, we observe that both weightings lead to roughly the same alpha over time. This pattern could indicate that the competition in the fixed income segment remains low, regardless of fund size, due to the higher complexity of the product. In addition, Garleanu and Pedersen (2018) argue that small investors tend to underperform because of their higher search costs and fees, while large investors are expected to outperform after a certain size because of their economies of scale and lower fees. For equity funds, we can confirm their theory. We observe higher aggregated alphas for institutional funds after costs, but similar alphas for retail equity managers before costs. Since retail funds can pool the investments of small investors, and mutual funds often manage retail and institutional money in the same aggregated fund, we expect this pattern. However, for fixed income funds, we find that retail funds achieve a much lower alpha after fees than their institutional competitors.

What is surprising in Figure 2, however, is the existence of three major periods where active managers underperformed their index counterparts: before the burst of the dot-com bubble, in the financial crisis, and in the recent past. Since we would have expected that active management pays in turbulent times, we will further explore this observation in the next section.

5.2 Drivers for the difference in performance of active and index funds

To gain further intuition about what drives a wedge between the performance of the average active fund and that of the average index fund, we ask whether the multi-factor model of Section 4.1 provides

some explanation for the difference in returns between the value-weighted portfolios of active and index mutual funds. Alerted by our observation from Figure 2, we enrich our regressions with the volatility index (VIX) of the Chicago Board Options Exchange (CBOE) as a fear gauge to proxy for market uncertainty.[33]

Table 6 shows the results for both active equity and fixed income funds when measured against the investable benchmark. Overall, the average values of R^2 absorb a significant fraction of the variance of the alpha before fees, in particular for fixed income funds and US institutional equity funds. For the equity funds in Panel A, we observe that the gross alpha loads profoundly and significantly on the SMB factor. Also, especially in the US, the performance difference loads negatively on the HML factor. The exposure is more pronounced for retail funds. Institutional funds, in contrast, have a much lower exposure to growth stocks. Overall, active funds seem to have a prominent small-cap bias and favor growth over value stocks. Furthermore, they tend to load positively on the momentum factor and negatively on the betting-against-beta factor.

[Table 6 about here.]

Concerning the VIX, we find that the difference in performance of active and index investing shows a negative sensitivity to changes in the VIX, which is often statistically significant. At first sight, this finding seems to run against our intuition, as we would expect active managers to use their skill to anticipate sudden uncertainty shocks. However, active managers that protect their portfolio against adverse shocks must pay an insurance premium in the long term. Such protection would generate relative losses to the market return in good times. Therefore, our result suggests that active managers prefer to run a short exposure to general market volatility, i.e., they tend to prefer small gains by selling insurance.

In Table 6, Panel B, we see that fixed income managers have a negative exposure to the shift factor. Consequently, they are less affected by rising interest rates. In exchange, they load on other risk factors to compensate for the lower expected returns. In particular, they load significantly on

[33] We downloaded the time-series of the VIX index of the Chicago Board Options Exchange (CBOE) from Bloomberg. In unreported results, we find that a high level of proxied uncertainty, e.g., by earnings-per-share volatility or dispersion of returns within a fund category, is in general favorable for the performance of actively managed funds. However, these effects are not significant.

the credit risk factor. As in Panel A for equity funds, fixed income managers also have a negative exposure to changes in the VIX. Therefore, they lose money if the VIX increases sharply, as it did during the latest financial crises, which also explains the large drop in the cumulative alpha in Figure 2 towards the end of 2008.

5.3 Persistence analysis

From an investor's perspective, it may be disappointing that active fund investing is, by and large, a zero-sum game after costs. How then can a fund investor do better and profit from actively managed funds? An initial idea is provided by Carhart (1997). He finds that US equity mutual funds with a substantial underperformance over the past year persist to underperform over the next year relative to a multi-factor benchmark. In contrast to the outperformance of the best mutual funds, he cannot explain the persistence in the worst mutual funds.[34] Thus, it would be of interest to know whether a fund investor is better-off if avoiding the losers of the past year.

To simulate the returns to an average active investor who trades according to this simple rule, we build momentum portfolios of active funds as follows. Every year in December, we first sort the active funds within each investment category based on their t-value for the value-weighted alpha measured against the investable benchmark.[35] Then, we invest in the value-weighted portfolio of the $x\%$ best performing active funds and normalize the weights each month. We repeat the same exercise for the $x\%$ worst performing funds. If one month there is no data for a particular fund, it disappears from the portfolio. To aggregate the performance numbers of the different investment categories, we value-weight the net returns by the total active assets.[36] We assume that funds do not charge transaction costs for incoming and outcoming investors.

[Table 7 about here.]

[34] Among others, Huij and Derwall (2008) find persistence in the US fixed income mutual funds market.

[35] To compute the alpha, we require at least ten of the twelve most recent monthly returns. When we sort by the t-value for the alpha, we consider the market risk of the fund and look at both the relative performance and the consistency of the relative performance against the benchmark.

[36] For $x\%$ we chose steps of 10% starting with all mutual funds to the best 10% mutual funds. We disregard data points where we have less than ten active mutual funds, and to calculate the benchmark return for the alpha we must have at least one index fund within the category for the past twelve months to start the out-of-sample backtest. For some small investment categories, over the year the number of funds drops below ten. In this case, we apply the next less restrictive filter.

Table 7 presents the momentum portfolio returns when selecting the best performing (Panel A) and the worst performing funds (Panel B). In Panel A, we find that the performance increases the more we focus on the best performing funds. For equity funds, the alpha after fees climbs from -0.23% (-0.60%) to a remarkable 0.62% (-0.02%) for institutional (retail) funds. For retail fixed income funds, alpha increases from -0.75% to -0.40%, but decreases for institutional fixed income funds from 0.26 to 0.16.[37] However, all these results are statistically insignificant, as the robust p-values remain high or even increase. Hence, after fees, even the best performing funds constitute a zero-sum game for the fund investor. This absence of persistence supports the theoretical argument of Berk and Green (2004) that persistence should not exist since new money flows into well-performing funds and there are diseconomies of scale, or because successful funds capture excess returns by raising fees.[38]

In Panel B of Table 7, we form portfolios by selecting the $x\%$ worst performing funds. We find that the value-weighted performance decreases drastically. For instance, for institutional equity funds it drops from -0.23% to -0.94%. However, only for retail funds does the negative performance of the 10% worst performing funds survive our robust alpha test adjusted for multiple hypothesis testing. For equity retail funds, the performance drops to -1.39% at the 5% significance level and, for fixed income retail funds, it drops to -0.84% at the 10% significance level. Hence, while Carhart (1997) shows for US equity funds and under a multi-factor benchmark that the persistence is significant for the worst performing funds, we can confirm this result only for retail equity and fixed income funds. Investing in these funds, obviously, is not a zero-sum game after costs. For institutional funds, we do not find such evidence. At the same time, Table 7 confirms Carhart (1997) in that we do not find any unexplainable persistence in overperforming funds.

Chen et al. (2004) and Yan (2008) find a negative relation between alpha and size and a positive relation with past return. Thus, on average, they find that future alpha is smaller for large funds but past returns are associated with higher future alpha, and predictability exists. Recently, Elton

[37]Interestingly, for fixed income institutional funds, we observe substantially lower betas the more we exclude badly performing funds from the portfolio, indicating that the best performing fixed income funds run a slightly different exposure than their investable benchmark suggests.

[38]As an additional exercise, we also explored the persistence of gross alphas. In unreported results, we find that the alphas increase substantially, but still they do not survive our statistical test, not even for the 10% best performing funds.

et al. (2012) show that alpha persistence does not disappear for larger funds. To explore the interplay between size and predictability, we perform a bivariate sort on size and persistence.

[Table 8 about here.]

Table 8 presents the alphas after fees for the different portfolios sorted according to fund size and the previous year's performance. We find that, except for fixed income retail funds, small winner funds perform better than their larger counterparts. They produce the highest alphas relative to their investable benchmarks. However, this outperformance is statistically insignificant. At the same time, we identify the small loser funds as the funds with the worst performance. For retail equity and fixed income funds, the negative alphas of small (and medium) loser funds become even statistically significant. Unsuccessful and small funds will continue to be unsuccessful, and they do so in a statistically significant way, while large funds tend to underperform but not significantly so. These findings resonate well with Berk and Green (2004)'s hypothesis that fund performance is inversely related to size due to diseconomies of scale.

5.4 Impact of fees

For US equity funds, Gil-Bazo and Ruiz-Verdú (2009) find a puzzling underperformance of mutual funds that charge higher fees. Their finding contradicts the argument of Habib and Johnsen (2016) that higher fees act as a signal for the unobservable quality of the costly research by active managers. However, higher fees can also be seen as a sure loss for investors, since they directly reduce the portfolio return when the quality of the manager is unobservable. Thus, higher fees imply lower net returns if the costly research of the active manager does not improve performance. To shed more light on this debate, we explore the impact of fees by proceeding analogously to the persistence analysis of the previous section.

[Table 9 about here.]

Using an investable benchmark, Table 9 presents the performance of active fund portfolios that include the $x\%$ least expensive (Panel A) and the $x\%$ most expensive (Panel B) funds of the preceding

year. In Panel A, the alpha against the investable benchmark increases for institutional equity funds from −0.23% to 0.83% and from −0.60% (−0.75%) to 0.30% (−0.40%) for retail equity (fixed income) funds. Hence, even the alpha of retail equity funds gets into positive territory if we exclude those funds that charge the highest fees. Again, although a fund investor can improve the performance in their fund portfolio by including only the least expensive funds, these improvements are statistically insignificant. For institutional fixed income funds, the alpha first increases but then decreases for the fund portfolio with the 20% and 10% lowest fees. Interestingly, the betas of these portfolios are substantially below one.

Panel B of Table 9 shows the performance of the portfolios with the x% most expensive funds. If the argument of Habib and Johnsen (2016) were valid and high fees were a signal of quality, we would expect increasing alphas the more we filter out the cheaper funds. However, we observe the opposite. The portfolios with the 10% most expensive funds perform poorly. For instance, the performance of the portfolio of the 10% institutional equity funds drops to −1.14%, compared to the 10% cheapest fund with an alpha of 0.83%. For equity retail funds, the underperformance of high-fee funds becomes significant already when we look at the 90% most expensive funds. For the 10% most expensive retail funds, we get a highly significant alpha of –2.83%, compared to the 10% least expensive retail funds with an alpha of 0.30%. Hence, if we only consider the universe of the most expensive equity retail funds, active investing definitely is no zero-sum game.

[Table 10 about here.]

Given the evidence that past winners and low fee funds generate a higher average alpha over time, we next ask whether the same pattern emerges when we control for performance persistence and fees simultaneously. Hence, we build nine portfolios that arise from the bivariate sort and the 30th and 70th percentiles for each criterion.[39] Table 10 shows the results. All active alphas are positive for the low-fee and winner portfolios except for the fixed income retail investor. Furthermore, for all sorted portfolios, the alpha in the top right (low fee and winner) corner is always larger than the alpha in the lower left (high fee and loser) corner. These alphas are insignificant, except for the equity retail funds. Here, the high-fee and loser portfolio has a highly significant negative alpha. Overall, we find

[39]For some small investment categories, there are time periods where none of the mutual funds belong to a particular group. In such a case, we invest in the value-weighted portfolio of all active funds within this category.

that both high performance and low fees over the past year have a positive impact on the alpha in the next year. The result is robust in the bivariate sort of the two criteria. Especially for retail funds, an investor is well advised to avoid high-fee loser funds as the negative alphas are highly significant under our testing framework.

[Figure 3 about here.]

To shed further light on which funds charge higher fees, Figure 3 shows the average active fees of the highly competitive US equity market over time together with the relative share of index funds in terms of assets under management. As expected, we find a substantial difference between the fees charged by retail and institutional funds, depending on the fund's age. Young retail funds charge the highest fees. However, the size of their fees has drastically decreased since the recent financial crisis, converging to the level of the fees charged by older retail funds. Interestingly, over the whole period, young institutional funds have charged fees similar to their older competitors. The gap between old retail and old institutional funds has been somewhat steady over the years, slightly narrowing recently. We also find that expense ratios are lower for large funds.[40] As Figure 3 suggests, over the years, the level of fees for active funds has tended to decrease further. Thus, we find further evidence for the zero-sum game hypothesis of Berk and Green (2004) in the sense that active managers start to adjust their fees due to the unabated growth of index funds.

5.5 How different are the results without the robust alpha test?

In our preliminary analysis, we demonstrated that fund returns are serially dependent. Furthermore, the simulation study in Appendix C indicated that our robust alpha test based block resampling is still liberal, but closer to the nominal level of the test. Consequently, we expect that the additional discoveries under the alternative test statistics are false positives. Still, it is natural to ask how our test, applied to our fund data, compares to other tests if we ignore serial dependence, and whether it would make a material difference to our empirical results.

[40] By the end of 2016, we find that the average expense ratio of the equal-weighted portfolio compared to the value-weighted portfolio is 21% higher for institutional equity funds, 37% higher for retail equity funds, 17% higher for institutional fixed-income funds, and 34% higher for retail fixed-income funds. Since institutional funds are usually larger than retail funds, this gap is consistent with the finding of Elton et al. (2012) that expense ratios are lower for larger funds.

We find that for the analysis of a single fund's performance compared to the investable benchmark, and in the single hypothesis setting, our robust alpha test decreases the proportion of unskilled (skilled) funds from 11.2% (5.4%) for the standard test to 9.0% (4.3%) at the 5% significance level. Hence, using the robust block-sampling test decreases the proportion of nonzero-alpha funds by almost 20%, compared to the standard test. For the fund portfolios, the decrease is almost 31%.

[Table 11 about here.]

For multiple hypothesis testing using FDR for single mutual funds, the impact of using our robust test is substantial. The decrease in nonzero-alpha funds when using our robust alpha test is 69% compared to the standard test and 60% compared to the standard resampling test at the 5% significance level. At the 10% significance level, the proportion of unskilled funds drops from 7% to 3.8% and from 1.6% to 0.8%, which corresponds to a decrease of nonzero-alpha funds of roughly 50%. Similarly, for the FWER applied to fund portfolios, we get a reduction of 27% when we use block resampling to take into account serial dependence. These differences are substantial and represent false discoveries due to the ignoring of serial dependence by the standard resampling. Admittedly, the proportions of skilled and unskilled funds are small. However, for an investor selecting a specific fund investment, discriminating between a skilled and unskilled fund or between fund portfolios with zero and nonzero alphas is absolutely crucial. Using our test statistics, we adequately take into account two prominent features of the data: serial and cross-sectional dependence. Thereby, we avoid potential distortions in the test statistics.

6 Conclusion

Analyzing a rich dataset from Morningstar covering 61,269 mutual funds from different regions and asset classes from 1992 to 2016 and comparing their returns to those from common multi-factor models, we find that a large fraction of active equity managers show zero alphas after fees. However, when we conduct a fair performance evaluation for equity and fixed income mutual funds, we find significant negative alphas after fees only for the "US Equity Large Cap Blend" for institutional funds and "Canada Fixed Income" for retail funds. For the vast majority of categories, we cannot reject

the hypothesis that active investing constitutes a zero-sum game after costs. Indeed, we even find categories such as "US Fixed Income" and "Global Equity Large Cap" for institutional investors with significant p-values before fees.

At first glance one would expect active managers to invest more carefully and take fewer risks. We have confirmed this hypothesis by the fact that active management takes a more conservative position with respect to the traditional risk factors, such as market and duration risk. However, we find that active equity and fixed income mutual funds are affected by adverse volatility shocks, suggesting that active managers sell protection in order to collect the insurance premium. Also, we find that, averaging over the different regions, the active investor has a higher sensitivity than index funds to alternative risk premia such as small cap and credit risk.

Sorting active fund portfolios according to their performance persistence, fees, and size, we find that low-fee winner portfolios and small winner portfolios tend to outperform but their alpha does not survive our test statistics. These results give further support to the zero-sum game argument of Berk and Green (2004). Our analysis also highlights some substantial differences between institutional and retail funds. In particular, our empirical results suggest to active retail investors that they should avoid avoid high-fee losers and small losers. Their alphas are negative and statistically significant, surviving our robust test statistics adjusted for multiple hypotheses.

References

Agarwal, Vikas, and Narayan Y. Naik, 2004, Risks and portfolio decisions involving hedge funds, *The Review of Financial Studies* 17, 63–98.

Bajgrowicz, Pierre, and Olivier Scaillet, 2012, Technical trading revisited: False discoveries, persistence tests, and transaction costs, *Journal of Financial Economics* 106, 473–491.

Banegas, Ayelen, Ben Gillen, Allan Timmermann, and Russ Wermers, 2013, The cross section of conditional mutual fund performance in European stock markets, *Journal of Financial Economics* 108, 699–726.

Barras, Laurent, Olivier Scalliet, and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance* 65, 179–216.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289–300.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* 29, 1165–1188.

Berk, Jonathan B., and Richard C. Green, 2004, Mutual fund flows and performance in rational markets, *Journal of Political Economy* 112, 1269–1295.

Berk, Jonathan B., and Jules H. van Binsbergen, 2015, Measuring skill in the mutual fund industry, *Journal of Financial Economics* 118, 1–20.

Berk, Jonathan B., and Jules H. van Binsbergen, 2017, Mutual funds in equilibrium, *Annual Review of Financial Economics* 9, 147–167.

Bonferroni, Carlo E., 1936, *Teoria statistica delle classi e calcolo delle probabilita* (Libreria Internazionale Seeber).

Breusch, Trevor S., and Adrian R. Pagan, 1980, The Lagrange multiplier test and its applications to model specification in econometrics, *Review of Economic Studies* 47, 239–253.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Chen, Joseph, Harrison Hong, Ming Huang, and Jeffrey D. Kubik, 2004, Does fund size erode mutual fund performance? The role of liquidity and organization, *American Economic Review* 94, 1276–1302.

Chen, Yong, Michael Cliff, and Haibei Zhao, 2017, Hedge funds: The good, the bad, and the lucky, *Journal of Financial and Quantitative Analysis* 52, 1081–1109.

DeMond, Andrew, Erdem Ultanir, and John Fox, 2012, Barra term structure models, Technical report, MSCI Inc.

Elton, Edwin J., 2001, A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases, *Journal of Finance* 56, 2415–2430.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake, 2012, Does mutual fund size matter? The relationship between size and performance, *The Review of Asset Pricing Studies* 2, 31–55.

Fama, Eugene F., and Kenneth R. French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427.

Fama, Eugene F., and Kenneth R. French, 2010, Luck versus skill in the cross-section of mutual fund returns, *Journal of Finance* 65, 1915–1947.

Ferreira, Miguel A., Aneel Keswani, António F. Miguel, and Sofia B. Ramos, 2013, The determinants of mutual fund performance: A cross-country study, *Review of Finance* 17, 483–525.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

French, Kenneth R., 2008, Presidential address: The cost of active investing, *Journal of Finance* 63, 1537–1573.

Garleanu, Nicolae B., and Lasse H. Pedersen, 2018, Efficiently inefficient markets for assets and asset management, *Journal of Finance* forthcoming.

Genest, Christian, and Bruno Rémillard, 2004, Test of independence and randomness based on the empirical copula process, *Test* 13, 335–369.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, 2015, Money doctors, *Journal of Finance* 70, 91–114.

Gerakos, Joseph, Juhani T. Linnainmaa, and Adair Morse, 2016, Asset managers: Institutional performance and smart betas, NBER Working Papers 22982, National Bureau of Economic Research.

Gil-Bazo, Javier, and Pablo Ruiz-Verdú, 2009, The relation between price and performance in the mutual fund industry, *The Journal of Finance* 64, 2153–2183.

Habib, Michel A, and D Bruce Johnsen, 2016, The quality-assuring role of mutual fund advisory fees, *International Review of Law and Economics* 46, 1–19.

Haerdle, Wolfgang, Joel Horowitz, and Jens-Peter Kreiss, 2003, Bootstrap methods for time series, *International Statistical Review* 71, 435–459.

Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 65–70.

Huij, Joop, and Jeroen Derwall, 2008, “Hot hands” in bond funds, *Journal of Banking and Finance* 32, 559–572.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65.

Kosowski, Robert, 2011, Do mutual funds perform when it matters most to investors? US mutual fund performance and risk in recessions and expansions, *The Quarterly Journal of Finance* 1, 607–664.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can mutual fund “stars” really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance* 61, 2551–2595.

Kuensch, Hans Rudolf, and Friedrich Goetze, 1996, Second-order correctness of the blockwise bootstrap for stationary observations, *The Annals of Statistics* 24, 1914–1933.

Lahiri, Soumendra N., 2003, *Resampling Methods for Dependent Data* (Springer–Verlag).

Ledoit, Oliver, and Michael Wolf, 2008, Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* 15, 850–859.

Ledoit, Olivier, and Michael Wolf, 2011, Robust performances hypothesis testing with the variance, *Wilmott* 2011, 86–89.

Moskowitz, Tobias J, 2000, Discussion, *The Journal of Finance* 55, 1695–1703.

Newey, Whitney K., and Kenneth D. West, 1994, Automatic lag selection in covariance matrix estimation, *The Review of Economic Studies* 61, 631–653.

Pastor, Lubos, Robert F. Stambaugh, and Lucian A. Taylor, 2015, Scale and skill in active management, *Journal of Financial Economics* 116, 23 – 45.

Patton, Andrew, Dimitris N. Politis, and Halbert White, 2009, Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White, *Econometric Reviews* 28, 372–375.

Pesaran, Hashem M., 2004, General diagnostic tests for cross section dependence in panels, CESifo Working Paper Series 1229, CESifo Group, Munich, Germany.

Politis, Dimitris N., and Halbert White, 2004, Automatic block-length selection for the dependent bootstrap, *Econometric Reviews* 23, 53–70.

Ratcliff, John W., and David E. Metzener, 1988, Pattern-matching: the gestalt approach, *Dr Dobbs Journal* 13, 46.

Romano, Joseph P., and Michael Wolf, 2005a, Exact and approximate stepdown methods for multiple hypothesis testing, *Journal of the American Statistical Association* 100, 94–108.

Romano, Joseph P., and Michael Wolf, 2005b, Stepwise multiple testing as formalized data snooping, *Econometrica* 73, 1237–1282.

Romano, Joseph P., and Michael Wolf, 2016, Efficient computation of adjusted p-values for resampling-based stepdown multiple testing, *Statistics & Probability Letters* 113, 38–40.

Sharpe, William F., 1991, The arithmetic of active management, *Financial Analysts Journal* 47, 7–9.

Wermers, Russ, 1999, Mutual fund herding and the impact on stock prices, *The Journal of Finance* 54, 581–622.

Yan, Xuemin Sterling, 2008, Liquidity, investment style, and the relation between fund size and fund performance, *Journal of Financial and Quantitative Analysis* 43, 741–767.

Zhang, Huazhu, and Cheng Yan, 2018, A skeptical appraisal of the bootstrap approach in fund performance evaluation, *Financial Markets, Institutions & Instruments* 27, 49–86.

A Description of the data

We summarize the steps for the data cleaning of the Morningstar database and provide summary statistics for the different asset classes and investment categories.

A.1 Raw Morningstar data

Our mutual fund sample is from the Morningstar database.[41] We focus on all funds with an *Investment Type* flagged by "Open-End Fund" or "Exchange-Traded Fund" including non-survivors from December 1991 to December 2016. We downloaded the following fields for each share class.

For the description of a share class, we retrieved the *Name*, *ISIN*, and *Base Currency*. It is common to name a share class starting with the name of the asset manager, followed by a description of the strategy, and an ending for the share class. For example, for the equity fund "Blackrock S&P 500 Index," there is a share class "Blackrock S&P 500 Index Institutional" for institutional and the "Blackrock S&P 500 Index Investor A" for retail clients.

The most specific categorization in Morningstar is the *Morningstar Category*, which is derived by analyzing the underlying portfolio holdings. In all, we find 504 different groups for the retail equity and fixed income funds. The *Global Category* combines several Morningstar categories, and we see a total of 68 groups for retail equity and fixed income funds. For example, the *Global Category* category "Europe Equity Large Cap" includes Morningstar categories, such as "EAA Fund Europe Large-Cap Blend Equity," "EAA Fund Europe Large-Cap Value Equity," "EAA Fund Europe Large-Cap Growth Equity," but also "US Fund Europe Stock" or "Canada Fund European Equity." Since we have within this broader categorization a higher chance of finding both index and equity funds, we concentrate on the *Global Category*. The *Global Broad Category Group* further aggregates the *Global Category* into the major asset classes. Since we focus on the comparison of active and index funds, we concentrate on the *Global Broad Category Group* "Equity" and "Fixed Income" funds. We thereby

[41] Recent work in Kosowski et al. (2006); Fama and French (2010); Barras et al. (2010) concentrates mostly on the survivor-bias-free CRSP US Mutual Fund Database. As shown by Elton (2001), the CRSP database also suffers from a survivorship bias: the so-called omission bias. Berk and van Binsbergen (2015) find that neither the CRSP nor the Morningstar database are free from errors. Thus, we must be careful, and we find the same errors as reported in this previous paper.

disregard categories such as "Allocation," "Money Market," or "Commodities" because for them there are insufficiently many index funds to make a fair comparison.

For the computation of the returns, we downloaded the following fields for each fund: *Monthly Return USD*, *Monthly Gross Return USD*, and *Net Assets - share class (Monthly) USD*. The *Monthly Return USD* includes management, administrative, and other costs that are deducted from the NAV, such as the 12b-1 fee. All income and capital gains are reinvested monthly. The *Monthly Gross Return USD* is based on the *Monthly Return USD* and adds the most recent net expense ratio. The *Net Assets - share class (Monthly) USD* is the monthly total net assets of a share class.

To distinguish between active and index funds, we make use of the *Index Fund* field. Those funds that track a particular index based on full replication or based on a representative sampling are flagged by Morningstar as index funds. Next, to filter the institutional and retail funds, we downloaded the field *Institutional*, which defines any fund as institutional if it either says "institutional" in the name of its share class, has a minimum investment above USD 100,000, or the prospectus says that it is for institutional investors only.

A.2 Data cleaning

For each fund, we retrieved its monthly net return, gross return, and total net assets, all in US dollars. We only included an observation if all three items were available. Often, and as reported in Berk and van Binsbergen (2015), we observe that net assets are reported quarterly or are missing for a specific month. In this case, we roll the assets under the assumptions of zero net flows, so as to increase the available data points and avoid disconnected time-series. Besides, for some institutional mutual funds, we observe zero fees because they are paid in separate contracts with the asset manager. Thus, we only include funds where the sum of the gross returns is larger than the sum of the net returns to exclude zero-fees funds. To avoid the incubation bias, we include funds only if they reach 5 million December 2016 US dollars in AUM.

We also see conversion errors, where funds assets suddenly increase by a high factor and then decrease again by a similar factor. First, we observe this behavior in emerging market currencies before 1999. Thus, we concentrate in the period before 1999 only on the developed currencies, Pound

Sterling, US Dollar, Euro, Singapore Dollar, Australian Dollar, Swedish Krona, South African Rand, Swiss Franc, Japanese Yen, New Zealand Dollar, Canadian Dollar, Norwegian Krone, Danish Krone. Also, we see that for some funds, the assets change by a factor higher than 100 and decrease in the next period to the same level as before the outlier. For these cases, we smooth the net assets over time if we see that the assets change by a factor higher than 10 and we decrease them in the next two periods by a factor of more than 0.5. But there are funds where this increase is verified by attaining the same fund levels in the future. Therefore, we only correct the assets if the same level is not exceeded in its future assets.

We also delete obvious mistakes, such as when an index fund shows high fees in the past and suddenly changes to a low fee. In this case, we keep only the low fee period, since we interpret this as being that either the fees were not correct or the fund changed from active to index. When we build the value-weighted portfolio for the investment categories, we also remove funds that show a beta below 0.05 relative to the average return of all the funds within the same investment category. Because of the low sensitivity to the average fund, these funds are not following a strategy similar to that of the rest of the group.

A.3 Aggregation of the share classes

Each line in the Morningstar dataset corresponds to a share class. In all, we obtain 435,453 lines of different share classes. Thus, we must aggregate the same share classes to avoid multiple tries of investment strategies by the same provider. First, we tried to use the fields *Administrator* and *Ticker of Fund's Oldest Share Class*; however, they are often missing. For this reason, we aggregated alphabetically subsequent mutual funds that are in the same Morningstar Category with the corresponding *Index Fund* flag and have a similar name. While Berk and van Binsbergen (2015) use the last word of the fund's name for the share class, we use the ratio provided by the SequenceMatcher of the difflib library in Python, which is based on the algorithm developed by Ratcliff and Metzener (1988) and, additionally, cleans the "junk" elements. We define two names to be similar if this ratio is above 0.8.

A.4 Summary statistics of investment categories

Table 1 shows the summary statistics of the cross-sectional monthly attributes across asset classes. Table A.1 provides a more detailed view of all the investment categories, where we find both index and active mutual funds. For the active funds, we analyze a total of 14,969 institutional and 46,300 retail funds, of which 56,136 are active funds and 5,133 are index funds. In general, there are fewer index funds, but they have higher average total net assets (TNA) and net returns, and also lower fees and about the same average number of years in the database. As expected, the institutional funds charge lower fees than their retail counterparts.

[Table A.1 about here.]

B The robust alpha test

Consider a fund with time-t return y_t and a set of K benchmark factor returns x_{tk}, $k = 1, ..., K$. A total of T returns are observed. We assume that these observations are generated by a stationary multivariate return distribution with mean vector μ and covariance matrix Σ:

$$\mu = \begin{pmatrix} \mu_y \\ \mu_{x_1} \\ \vdots \\ \mu_{x_K} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{yx_1} & \cdots & \sigma_{yx_K} \\ \sigma_{x_1y} & \sigma_{x_1}^2 & \cdots & \sigma_{x_1x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_Ky} & \sigma_{x_Kx_1}^2 & \cdots & \sigma_{x_K}^2 \end{pmatrix}, \tag{B.1}$$

with the observed means $\hat{\mu}$ and sample covariance matrix $\hat{\Sigma}$. By defining a vector $\mu_X = (0, E[x_1], \cdots, E[x_K])'$, we can express the fund's alpha as

$$\alpha = E[y] - \mu_X' \Sigma_{XX}^{-1} y_X, \tag{B.2}$$

with

$$\Sigma_{XX} = \begin{pmatrix} 1 & E[x_1] & E[x_2] & \cdots & E[x_K] \\ E[x_1] & E[x_1^2] & E[x_1x_2] & \cdots & E[x_1x_K] \\ \vdots & \vdots & & \ddots & \vdots \\ E[x_K] & E[x_Kx_1] & E[x_Kx_2] & \cdots & E[x_K^2] \end{pmatrix} \quad \text{and} \quad y_X = \begin{pmatrix} E[y] \\ E[x_1y] \\ \vdots \\ E[x_Ky] \end{pmatrix}. \tag{B.3}$$

Then, we test for the hypothesis

$$H_0: \quad \alpha = 0 \qquad \text{against} \qquad H_1: \quad \alpha \neq 0. \tag{B.4}$$

Furthermore, we define $\zeta_k = E[yx_k]$, $\gamma_k = E[x_k^2]$, $\xi_{kj} = E[x_kx_j]$, $j > k$, and the combined vector $\nu = (\mu_y, \cdots, \mu_{x_k}, \cdots, \zeta_k, \cdots, \gamma_k, \cdots, \xi_{kj}, \cdots)' \in \mathbb{R}^{1+3k+k(k-1)/2}$ with sample counterpart $\hat{\nu}$. Now, we can express the true alpha as a function f of ν:

$$\alpha = E[y] - \mu_X' \Sigma_{XX}^{-1} y_X = f(\nu); \tag{B.5}$$

and the estimated alpha as function of $\hat{\nu}$: $\hat{\alpha} = f(\hat{\nu})$. As mentioned in Ledoit and Wolf (2008), under mild regularity conditions,

$$\sqrt{T}(\hat{\nu} - \nu) \xrightarrow{d} N(0, \Psi), \tag{B.6}$$

where Ψ is an unknown symmetric positive semi-definite matrix. By the delta method, we obtain

$$\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \nabla' f(\nu) \Psi \nabla f(\nu)) \tag{B.7}$$

with

$$\nabla' f(\nu) = \left(\frac{\partial f(v)}{\partial \mu_y}, \cdots, \frac{\partial f(v)}{\partial \mu_{x_k}}, \cdots, \frac{\partial f(v)}{\partial \zeta_k}, \cdots, \frac{\partial f(v)}{\partial \gamma_k}, \cdots, \frac{\partial f(v)}{\partial \xi_{kj}}, \cdots \right).' \tag{B.8}$$

Given a consistent estimator $\hat{\Psi}$ of Ψ, we can compute a standard error for $\hat{\alpha}$ by

$$s(\hat{\alpha}) = \sqrt{\frac{\nabla' f(\nu) \hat{\Psi} \nabla f(\nu)}{T}}. \tag{B.9}$$

To test the null hypothesis in Equation (B.4), we focus on the bootstrap inference for time-series data outlined in Ledoit and Wolf (2008). In particular, we denote the optimal block length by b and define $l = \text{floor}(T/b)$. As shown in Kuensch and Goetze (1996), the bootstrapped estimator of $\hat{\Psi}^*$ is

$$\hat{\Psi}^* = \frac{1}{l}\sum_{j=1}^{l} \eta_j \eta_j', \tag{B.10}$$

where

$$\begin{aligned} z_t^* &= \left(y_t^* - \hat{\mu_y}^*, ..., x_{tk}^* - \hat{\mu_x}^*, ..., y_t x_{tk} - \hat{\zeta}_k^*, ..., {x_{tk}^*}^2 - \hat{\gamma}_k^*, ..., x_{tk}^* x_{tj}^* - \hat{\xi}_{kj}, ...\right), \\ \eta_j &= \frac{1}{\sqrt{b}}\sum_{t=1}^{b} z_{(j-1)b+t}. \end{aligned} \tag{B.11}$$

Next, the studentized statistics are

$$\tilde{d}_m^* = \frac{|\hat{\alpha}_m^* - \hat{\alpha}|}{s\left(\hat{\alpha}_m^*\right)}, \tag{B.12}$$

and the p-value is

$$PV = \frac{\{\tilde{d}_m^* \geq \hat{d}\} + 1}{M + 1}, \tag{B.13}$$

where $\hat{d}$ is the original studentized test statistic that was computed from the observed returns. We use Newey–West standard errors to calculate the original standard errors. Regarding the optimal block, we suggest using either the method of Politis and White (2004) and Patton et al. (2009) for the univariate case, or the method of Ledoit and Wolf (2008) for the bivariate case. For our empirical analysis, we would like to compare up to 30 different investment categories and, so far, there is no available method to make this comparison. Consequently, we will further discuss the optimal block size to use in our simulations in Appendix C.

C Accuracy of the robust alpha test

We now present the results of a simulation study to show the difference between our robust alpha test and the standard hypothesis tests.[42] For this purpose, we first simulate a single hypothesis setting.

[42] Other papers that extensively use bootstrap techniques often do not perform such a simulation study to validate their approach.

For realistic time-series, we select the first ten US mutual funds of the Morningstar database within the category "US Equity Large Cap Blend" that offer the entire return history from 1992 to 2016 (T=300). As benchmark models, we focus on the one-factor "CAPM," i.e., the market excess return, the three-factor "FF3," and the five-factor "FF5" model. For the data generating process (DGP), we sample from the realized returns with a circular block bootstrap and block sizes of 1, 3, and 6. We selected this grid of block sizes based on our analysis in Section 2, where we observe that most of the optimal circular block sizes range from one to six. This grid corresponds to time periods of one, three, and six months. The block sizes of three and six are the ones that take the evidence of serial dependence from Section 2 into account. A block size of one generates independent data, and we employ this block size only for reasons of comparison. For each fund, we simulate 1,000 paths and set the alpha under the null hypothesis to the true observed alpha of the data. The bootstrapped p-values (Boot) are then calculated as illustrated in Appendix B by employing M = 1,000 and the optimal block size by the method of Politis and White (2004) and the correction of Patton et al. (2009). We compare the robust p-values with those from the standard inference methods; that is, based on the normal distribution (Standard), Newey–West (NW), and HC3 standard errors.

[Table C.1 about here.]

Table C.1 shows the empirical rejection probabilities of the falsely rejected null hypothesis compared to the nominal levels α = 10%, α = 5%, and α = 1%. Because the null hypothesis is true for all the simulations, the true rejection probabilities should be equal to the nominal levels of the test. If a test shows a higher percentage of rejections, then we regard this test as too liberal. While we observe that the standard inference tests based on the normal distribution, the Newey–West, and HC3 standard errors are too liberal in rejecting the null hypothesis, the bootstrapped solution (Boot) presented in the previous section is close to the nominal levels. We highlight in bold the empirical rejection probabilities that are closest to the desired level. We observe the HC3 standard errors to be in some cases closer to the desired level than are those of the block bootstrapped method, but only in the case where we apply the standard but less realistic bootstrap with a block size of one where we lose any dependence over time. However, as we demonstrate in Section 2, the optimal block size, and thus a realistic assumption for the DGP is, in general, around three or six, for which our

bootstrapped test is tailored to be more accurate.[43]

Since there is still the open question of the optimal block size in the multiple hypothesis setting when controlling the FWER, as illustrated in Romano and Wolf (2005a,b, 2016), we conduct a second simulations study. Unlike the single mutual fund analysis, where we regard each fund in isolation and then apply the multiple hypothesis framework of Barras et al. (2010), in this case, we must consider the cross-dependence structure, and jointly sample the funds and benchmark returns. For this purpose, we focus on the 17 portfolios within the "Inv. Categories" setting from Section 2 with the investable one-factor benchmark model that is based on the value-weighted return of index funds. Also, instead of calculating the Type I Errors as in the single hypothesis setting, we compute the empirical rejection probabilities based on the FWER, as illustrated in Romano and Wolf (2005a,b, 2016). To find the optimal block size that is closest to the nominal levels of the test, we focus on the following grid of block sizes: 1, 3, 6, 9, and 12. Regarding the DGP, we keep the grid from our first simulation study.

[Table C.2 about here.]

Table C.2 shows the empirical rejection probabilities based on the FWER. Likewise, for the FWER, we find the bootstrapped robust alpha test to achieve the desired levels at optimal block sizes three or six. Given that for a block size of three we observe accurate rejection probabilities, in the multiple comparisons of portfolios, we will, in the remainder of the paper, present the results based on the optimal block size of three. Finally, the more conservative block sizes six and nine are applied for robustness checks.

[43] A similar observation was also made in Ledoit and Wolf (2008) for testing the Sharpe Ratio and in Ledoit and Wolf (2011) for the variance.

Figure 1: Value-weighted alpha of active mutual funds within investment categories

Multiple hypotheses adjusted p-value (y-axis) and annualized value-weighted alpha of active versus index funds (x-axis) for all investment categories as defined by the "Global Category" of Morningstar. Top (bottom): analysis after (before) management fees. We form the four groups with the combinations retail and institutional as well as the periods 1992–2016 and 2000–2016.

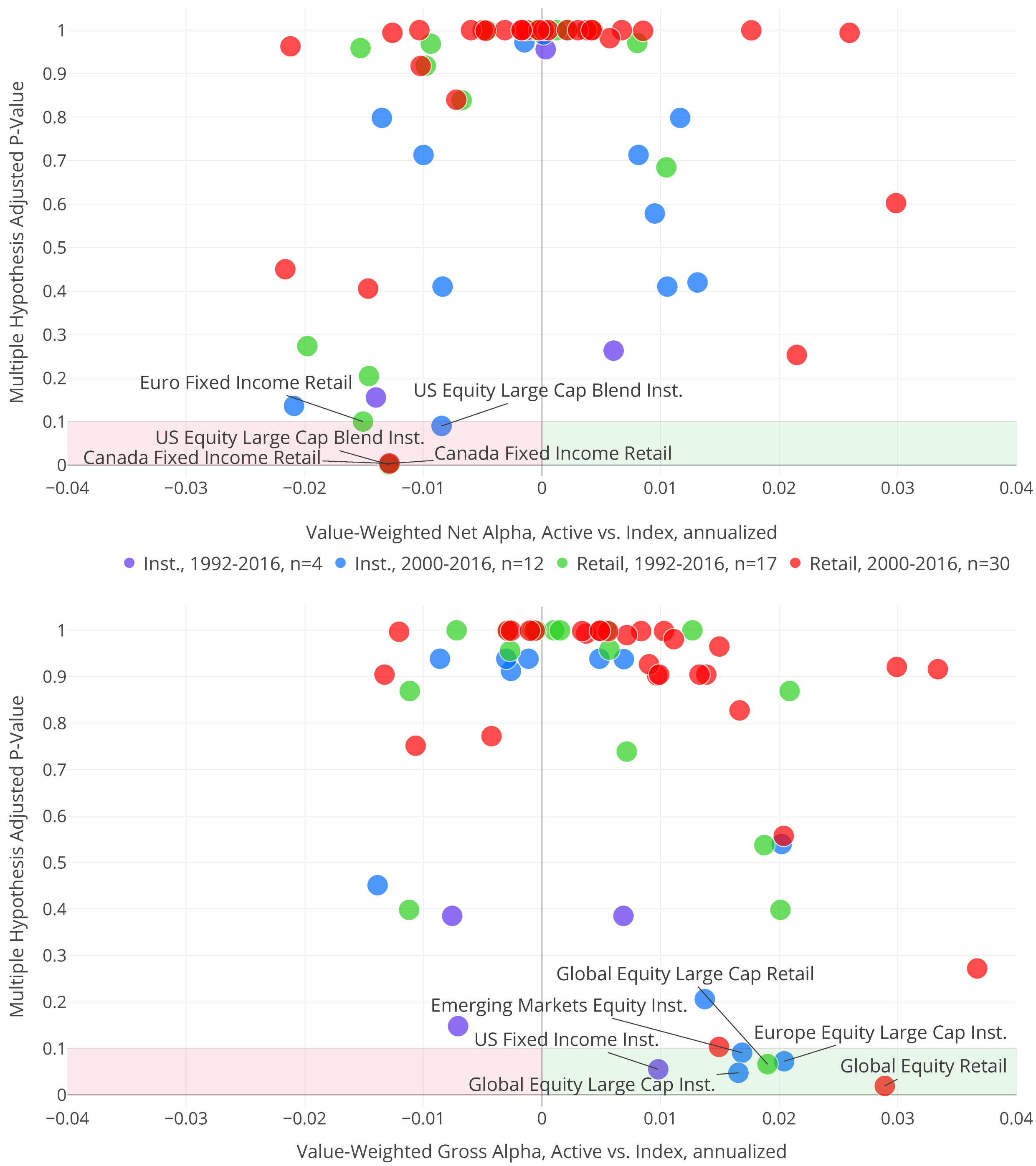

Figure 2: Aggregated value-weighted alpha of active minus index

Cumulated logarithmic alphas for the active equity (top) and fixed income (bottom) mutual funds. The alpha is the value-weighted return of the active funds against the value-weighted return of the index funds within the same investment category. The figure shows the aggregated alpha with equal-weights (EW) and value-weights (VW) across the Morningstar investment categories. We analyze both institutional and retail funds. We also regard the portfolios before (Gross) and after (Net) costs. We include all mutual funds within Morningstar where net and gross returns and assets under management are available, and where we have at least one index fund within the same investment category. The analysis is in US dollars.

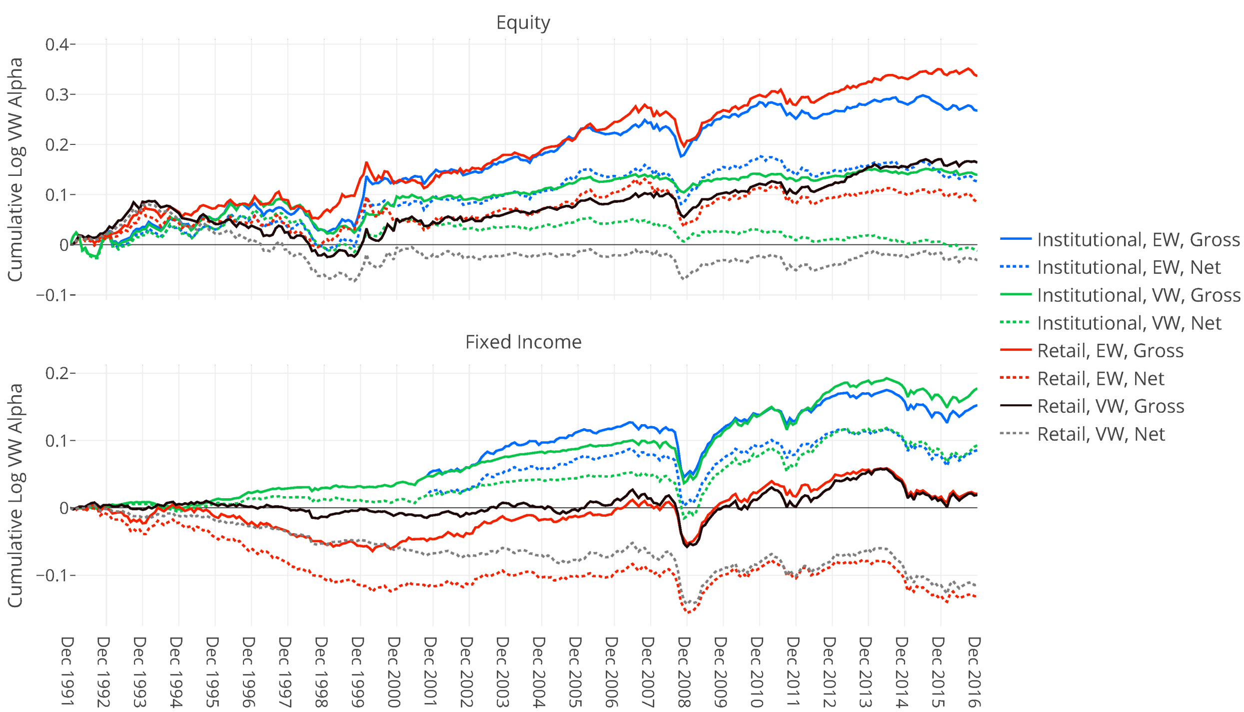

Figure 3: Active equity fees of young and old funds in the US

Average active fee over the last year (top) of US equity funds with a track record of more than five years (>5y) and with a track record of at most one year (<1y), and the percentage of index funds (bottom) within all US equity mutual funds. We distinguish between retail and institutional funds.

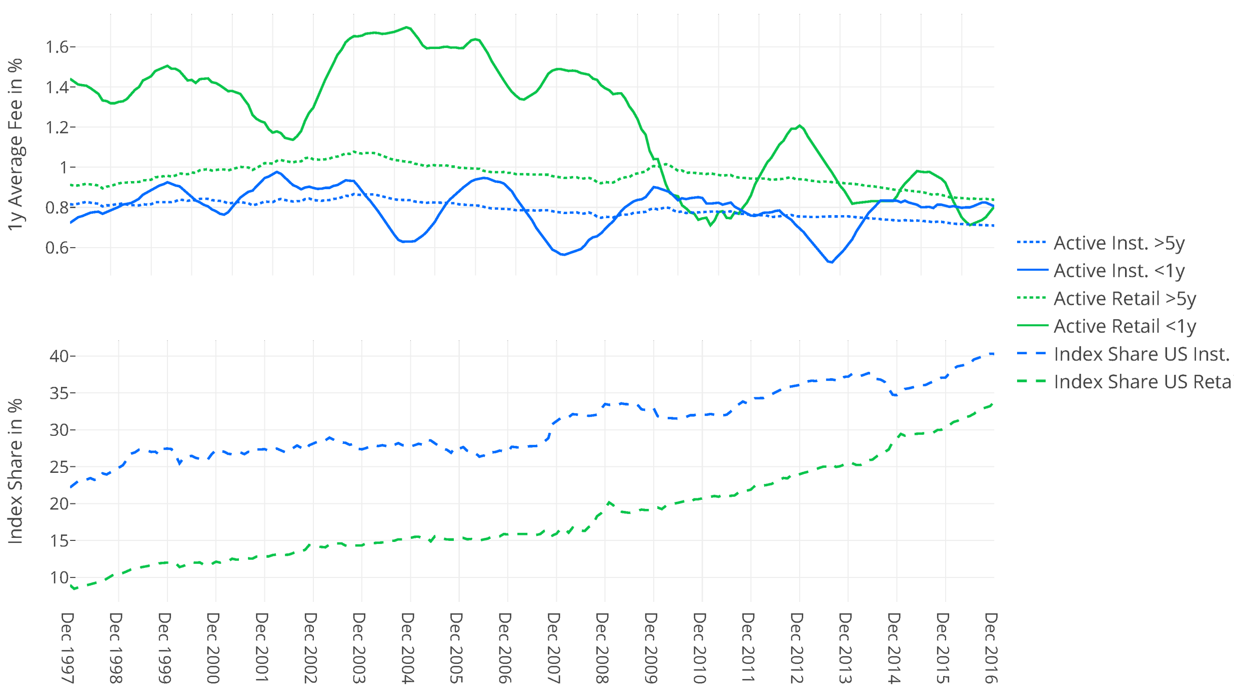

Table 1: Mutual fund database summary statistics

Total number, average number (Avg Number), the average total net assets in million USD (Avg TNA), average annual net return in USD (Avg Net Ret), average annual fee in USD (Avg Fees ann), and the average years of a fund in the database (Avg Years) over the time period from December 1991 to December 2016 of all available funds in the Morningstar database flagged by Open-End or Exchange-Traded funds. We only include funds within the "Global Broad Category Group" equity (Equity) and fixed income (Fixed Income) for which we provide the category statistics. The average corresponds to the mean of cross-sectional monthly attributes.

in USD		Total Number		Avg Number		Avg TNA		Avg Net Ret		Avg Fees ann		Avg Years	
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*
Equity	*Inst.*	8,488	691	2,506.1	199.5	255.3	910.0	8.85%	8.95%	0.86%	0.15%	7.4	7.2
	Retail	26,741	3,551	9,147.8	950.6	453.6	732.2	8.16%	8.27%	1.18%	0.31%	8.6	6.7
Fixed	*Inst.*	5,566	224	1,440.3	57.8	333.2	663.0	4.97%	5.19%	0.54%	0.19%	6.5	6.5
Income	*Retail*	15,341	667	4,545.6	152.6	346.7	817.6	4.87%	5.20%	0.88%	0.25%	7.4	5.7

Table 2: Dependence analysis

Results of dependence analysis. Panel A shows the total number of funds within each category and the percentage of mutual funds with a significant serial dependence according to the Ljung–Box (LJ) and Genest and Rémillard (2004) (GR) tests. For the latter, we use 1,000 simulations. Panel B shows the average correlations (AvgCorr) and the p-values for the cross-sectional dependence test of Pesaran (2004). We analyze the residuals of single mutual funds (Single) and portfolios of mutual funds (Portfolio), split into equity (Equity) and fixed income (FixedInc) mutual funds. For the portfolio of mutual funds, we additionally analyze all 63 investment category portfolios (InvCat) from Figure 1. For the benchmark model, we use the investable one-factor model with the value-weighted return of the index funds within the same category as the analyzed single mutual fund or portfolio of mutual funds (Inv), the equity five-factor model (FF5) with the regional factors "market," "size," and "value" of Fama and French (1992), and also "momentum" of Jegadeesh and Titman (1993) and "betting against beta" of Frazzini and Pedersen (2014), and the regional four-factor fixed income model (FI4) with the "shift," "twist" and "butterfly" factors, as well as the difference between the BBB and AAA credit spread.

Panel A: Serial dependence

		Single				Portfolio		
		Equity		FixedInc		InvCat	Equity	FixedInc
		Inv	FF5	Inv	FI4	Inv	FF5	FI4
Retail	Total	24,456	12,816	14,579	4,719	47	5	4
	GR<5%	17%	20%	22%	16%	28%	60%	25%
	LJ<5%	19%	23%	19%	18%	19%	80%	25%
Inst.	Total	7,025	3,817	4,815	1,528	16	5	4
	GR<5%	14%	15%	15%	18%	31%	0%	0%
	LJ<5%	15%	18%	19%	21%	44%	20%	25%

Panel B: Cross-sectional dependence

		Single				Portfolio		
		Equity		FixedInc		InvCat	Equity	FixedInc
		Inv	FF5	Inv	FI4	Inv	FF5	FI4
Retail	AvgCorr	0.04	0.22	0.11	0.62	0.14	0.11	0.44
	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Inst.	AvgCorr	0.06	0.21	0.15	0.60	0.12	0.06	0.41
	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Overview of the statistical methodology

Whether different statistical methodologies can cope with potential dependencies in the data, serial dependence (Time) and cross-sectional dependence (Cross). To obtain the p-values needed for applying the two multiple-hypothesis tests, the FDR of Barras et al. (2010) and the FWER of Romano and Wolf (2016), use either the normal distribution (Standard), Newey–West (NW), HC3, the standard resampling (Robust-SR), or the block resampling (Robust-BR) standard errors. For the NW standard errors we use the automatic bandwidth selection procedure described in Newey and West (1994) based on the Bartlett kernel. The HC3 standard errors are consistent even in the presence of heteroscedasticity of an unknown form. The last two columns show what the data tells us concerning the dependencies for single funds and for the fund portfolios.

	Dependence:				Dependence:	
	Serial	Cross-sectional			Serial	Cross-sectional
Methodology:		FDR	FWER	Evidence/Data:		
Standard:	×	✓	not applicable	Single funds:	✓	✓
NW:	✓	✓	not applicable	Portfolios:	✓	✓
HC3:	×	✓	not applicable			
Standard-RS:	×	✓	✓			
Block-RS:	✓	✓	✓			

Table 4: Single mutual equity and fixed income funds against multi-factor benchmark

Proportion (in percentages) of zero, positive, and negative alpha funds, and significant p-values based on the FDR at the 10% significance level for funds with a positive and negative alpha. The alpha is after fees and relative to a multi-factor benchmark and across the equity investment regions US, Global, Europe, Japan, and Asia ex-Japan, as well as for the USD, CHF, EUR, and GBP fixed income markets. The results are based on the method of Barras et al. (2010), while we apply our robust alpha test to compute the single mutual funds' p-values. We show the results for the five-factor equity benchmark model based on the regional MKT, SMB, HML, WML, and BAB factors. For the fixed income benchmark, we include the four local factors "shift," "twist," and "butterfly," as well as the AAA–BBB credit spread.

Panel A: Retail funds

		Equity - 5 factors						Fixed Income - 4 factors				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	55.1	39.6	66.2	67.9	83.0	62.3	38.9	71.0	77.8	83.3	67.8
	Positive alpha	0.0	0.0	3.0	5.7	0.0	1.7	23.3	3.3	22.2	16.7	16.4
	Negative alpha	44.9	60.4	30.8	26.4	17.0	35.9	37.8	25.7	0.0	0.0	15.9
	FDR 10 alpha>0	0.0	0.1	0.0	0.0	0.0	0.0	18.3	0.0	0.0	0.0	4.6
	FDR 10 alpha<0	30.9	64.9	16.5	15.2	0.0	25.5	36.9	0.9	0.0	0.0	9.4
Index	*Zero alpha*	61.9	30.1	76.5	73.7	100.0	68.4	41.6	93.3	71.6	100.0	76.6
	Positive alpha	0.0	0.0	3.6	5.9	0.0	1.9	29.2	6.7	28.4	0.0	16.1
	Negative alpha	38.1	69.9	19.9	20.4	0.0	29.7	29.2	0.0	0.0	0.0	7.3
	FDR 10 alpha>0	0.0	0.0	0.0	1.3	0.0	0.3	5.0	0.0	0.0	0.0	1.3
	FDR 10 alpha<0	25.3	79.6	0.0	6.6	0.0	22.3	15.0	0.0	0.0	0.0	3.8

Panel B: Institutional funds

		Equity - 5 factors						Fixed Income - 4 factors				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	69.3	53.5	78.5	88.2	97.4	77.4	38.5	77.4	60.1	82.7	64.7
	Positive alpha	0.0	0.0	8.2	9.4	0.0	3.5	40.7	22.6	39.9	17.3	30.1
	Negative alpha	30.7	46.5	13.3	2.4	2.6	19.1	20.8	0.0	0.0	0.0	5.2
	FDR 10 alpha>0	0.0	0.0	0.0	0.0	0.0	0.0	42.1	0.0	0.2	0.0	10.6
	FDR 10 alpha<0	3.0	38.0	0.6	0.5	0.0	8.4	18.1	0.0	0.0	0.0	4.5
Index	*Zero alpha*	66.9	55.7	91.9	92.5	90.6	79.5	57.5	71.4	56.5	95.0	70.1
	Positive alpha	0.0	0.0	6.8	0.0	0.0	1.4	21.3	26.2	43.5	0.0	22.7
	Negative alpha	33.1	44.3	1.4	7.5	9.4	19.1	21.3	2.4	0.0	5.0	7.2
	FDR 10 alpha>0	0.0	0.0	0.0	5.0	0.0	1.0	5.0	0.0	0.0	0.0	1.3
	FDR 10 alpha<0	17.2	37.1	0.0	0.0	0.0	10.9	15.0	0.0	0.0	0.0	3.8

Table 5: Single mutual equity and fixed income funds against investable benchmark

Proportion (in percentage numbers) of zero, positive, negative alpha funds, and significant p-values based on the FDR at the 10% significance level for funds with a positive and negative alpha. The alpha is after fees and relative to the investable value-weighted portfolio of index funds within the same Morningstar investment category. The results are based on the method of Barras et al. (2010), while single mutual funds' p-values are derived from our robust alpha test.

Panel A: Retail funds

		Equity - Investable						Fixed Income - Investable				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	71.9	78.3	78.0	87.3	94.3	82.0	63.3	58.2	53.1	95.9	70.5
	Positive alpha	0.0	12.2	8.6	0.0	0.0	4.2	19.1	0.0	0.0	0.0	4.7
	Negative alpha	28.1	9.5	13.4	12.7	5.7	13.9	17.6	41.8	46.9	4.1	24.9
	FDR 10 alpha>0	0.0	1.4	0.0	0.0	0.4	0.3	6.7	0.0	0.0	0.0	1.4
	FDR 10 alpha<0	6.3	0.8	2.1	2.0	0.0	2.2	8.3	7.4	35.3	0.0	10.6
Index	*Zero alpha*	64.2	62.8	88.0	100.0	100.0	83.0	58.2	89.0	73.8	97.1	80.2
	Positive alpha	2.5	30.3	4.1	0.0	0.0	7.4	21.2	0.0	11.1	0.0	7.9
	Negative alpha	33.3	6.9	7.9	0.0	0.0	9.6	20.7	11.0	15.1	2.9	11.8
	FDR 10 alpha>0	0.0	22.0	0.0	0.0	0.0	4.4	1.5	0.0	0.0	0.0	1.2
	FDR 10 alpha<0	26.1	1.7	0.0	0.7	0.0	5.7	2.0	2.4	0.0	0.0	2.0

Panel B: Institutional funds

		Equity - Investable						Fixed Income - Investable				
		US	*Global*	*Europe*	*Japan*	*Asia*	Avg.	*USD*	*CHF*	*EUR*	*GBP*	Avg.
Active	*Zero alpha*	80.1	81.2	83.5	89.6	100.0	86.9	55.2	75.4	60.1	100.0	75.5
	Positive alpha	0.0	14.3	13.3	2.7	0.0	6.0	33.5	0.0	0.0	0.0	7.9
	Negative alpha	19.9	4.5	3.2	7.7	0.0	7.1	11.4	24.6	39.9	0.0	16.6
	FDR 10 alpha>0	0.0	1.2	0.0	0.0	0.0	0.2	18.2	0.0	0.0	0.0	3.7
	FDR 10 alpha<0	0.2	0.3	0.0	0.0	0.4	0.2	4.9	0.0	13.2	0.0	3.7
Index	*Zero alpha*	53.1	44.6	92.9	100.0	94.4	77.0	95.6	90.8	78.6	100.0	88.4
	Positive alpha	6.1	44.1	7.1	0.0	2.8	12.0	2.2	0.0	0.0	0.0	2.8
	Negative alpha	40.8	11.4	0.0	0.0	2.8	11.0	2.2	9.2	21.4	0.0	8.8
	FDR 10 alpha>0	0.0	48.5	0.0	0.0	0.0	9.7	2.2	0.0	0.0	0.0	2.4
	FDR 10 alpha<0	34.6	2.0	0.0	0.0	0.0	7.3	0.0	0.0	0.0	0.0	1.5

Table 6: Performance drivers of active minus index

Results from regressing the difference between value-weighted active investing and index investing, before fees. For the benchmark model, we include the difference between the VIX index and the regional equity model with the regional MKT, SMB, HML, WML, and BAB factors. For the regional fixed income model, we add the VIX index to the four local factors shift, twist, and butterfly (BFLY), as well as the AAA–BBB credit spread (SPR). Coefficient estimates are multiplied by 100 and HC3 standard errors are in parentheses. By *, **, and *** we denote p-values below 0.1, 0.05, and 0.01, respectively. The last rows report the adjusted R^2 values.

Panel A: Equity funds

	Retail					Institutional				
	US	*Global*	*Europe*	*Japan*	*Asia ex-Japan*	*US*	*Global*	*Europe*	*Japan*	*Asia ex-Japan*
Const.	-0.01	0.27^{***}	0.04	0.10	-0.15	-0.06^{*}	0.15^{***}	0.07	0.07	0.17
	(0.04)	(0.06)	(0.05)	(0.09)	(0.20)	(0.03)	(0.05)	(0.07)	(0.07)	(0.15)
MKT	-2.41^{**}	-7.07^{***}	4.04^{**}	1.88	3.97	-0.25	-2.80^{**}	1.34	2.37	-4.71^{*}
	(1.09)	(1.90)	(1.60)	(2.56)	(6.05)	(1.08)	(1.28)	(1.90)	(1.62)	(2.49)
SMB	13.77^{***}	14.87^{***}	23.14^{***}	7.58^{**}	6.73	26.39^{***}	14.72^{***}	35.59^{***}	15.63^{***}	22.50^{***}
	(1.84)	(2.93)	(3.10)	(3.77)	(4.94)	(1.53)	(2.04)	(4.33)	(3.29)	(4.98)
HML	-6.93^{***}	-3.85	-8.32^{***}	-12.41^{***}	13.73^{***}	-3.95^{**}	-0.57	2.15	-1.19	22.73^{***}
	(1.74)	(2.69)	(2.29)	(3.27)	(4.95)	(1.77)	(2.25)	(3.09)	(3.07)	(5.42)
WML	2.73^{*}	4.78^{***}	2.75^{*}	12.47^{***}	-0.47	2.09^{*}	2.01	5.88^{***}	-1.54	0.20
	(1.42)	(1.61)	(1.57)	(3.26)	(4.25)	(1.26)	(1.28)	(1.83)	(2.46)	(3.79)
BAB	0.41	-9.50^{***}	-6.51^{***}	-3.00	5.51	0.39	0.28	-6.59^{***}	-8.60^{***}	-12.41^{**}
	(1.52)	(2.61)	(1.75)	(2.55)	(6.61)	(1.34)	(1.99)	(2.28)	(2.39)	(5.84)
$\Delta\,VIX$	-3.46^{***}	-7.16^{***}	-2.94^{*}	-3.38	7.63	-2.39^{**}	-3.57^{***}	-3.68^{**}	-3.48^{*}	-4.02
	(0.98)	(2.00)	(1.75)	(2.36)	(6.11)	(1.05)	(1.34)	(1.69)	(1.80)	(3.51)
R^2	0.51	0.20	0.26	0.25	0.08	0.73	0.18	0.45	0.22	0.21

Panel B: Fixed income funds

	Retail				Institutional			
	USD	*CHF*	*EUR*	*GBP*	*USD*	*CHF*	*EUR*	*GBP*
Const.	0.05^{***}	0.02	0.03^{*}	0.13^{**}	0.06^{***}	-0.00	0.02	0.08^{*}
	(0.01)	(0.02)	(0.02)	(0.05)	(0.02)	(0.01)	(0.03)	(0.04)
SHIFT	-0.91^{***}	-2.21^{***}	-2.73^{***}	-2.48^{***}	-0.71^{***}	-0.50^{***}	-2.26^{***}	-0.23
	(0.09)	(0.27)	(0.17)	(0.48)	(0.17)	(0.09)	(0.24)	(0.36)
TWIST	-0.33^{*}	-0.82	-0.09	-1.40	-1.33^{***}	-0.37^{**}	-0.27	1.11^{*}
	(0.17)	(0.57)	(0.26)	(1.45)	(0.29)	(0.16)	(0.45)	(0.59)
BFLY	-0.30	0.99^{*}	0.76	2.40	-0.72	-0.31	0.85	1.09
	(0.32)	(0.51)	(0.60)	(2.26)	(0.59)	(0.28)	(0.96)	(1.23)
SPR	1.27^{***}	0.30^{**}	0.27	3.81^{***}	1.13^{***}	0.12^{**}	-0.11	1.60^{***}
	(0.21)	(0.14)	(0.22)	(0.90)	(0.33)	(0.06)	(0.30)	(0.47)
$\Delta\,VIX$	-1.86^{**}	-2.97^{***}	-1.17	-5.34^{*}	-1.40	-0.17	-2.86^{**}	-0.90
	(0.73)	(1.01)	(0.81)	(2.91)	(1.19)	(0.24)	(1.20)	(1.78)
R^2	0.82	0.59	0.81	0.72	0.63	0.25	0.63	0.49

Table 7: One-year persistence of the alpha after fees

Annualized alpha after fees (in %), the corresponding block-bootstrapped multiple hypothesis adjusted p-value (in brackets), and the beta for the value-weighted performance of active mutual funds benchmarked against the value-weighted performance of index funds. In Panel A, each row from 100% to 10% corresponds to the value-weighted portfolio including only the x% best active mutual funds of the past year based on the t-value for the alpha. In Panel B, we report the same numbers but for the x% worst active mutual funds of the past year. Every December, we rebalance the momentum portfolios. The data sample ranges from 1993 to 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

Panel A: Value-weighted performance of the x% best performing funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.12	−0.12	−0.10	−0.01	0.07	0.07	0.11	0.40	0.62
			(0.78)	(0.94)	(0.94)	(0.95)	(0.98)	(0.97)	(0.97)	(0.97)	(0.73)	(0.65)
		beta	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98
	Retail	*alpha*	−0.60	−0.46	−0.44	−0.41	−0.44	−0.40	−0.31	−0.27	−0.13	−0.02
			(0.30)	(0.52)	(0.52)	(0.53)	(0.52)	(0.53)	(0.64)	(0.71)	(0.91)	(0.97)
		beta	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Fixed Income	Inst.	*alpha*	0.26	0.32	0.35	0.33	0.30	0.28	0.27	0.22	0.13	0.16
			(0.75)	(0.68)	(0.59)	(0.59)	(0.66)	(0.68)	(0.68)	(0.75)	(0.75)	(0.75)
		beta	0.88	0.87	0.86	0.86	0.85	0.85	0.85	0.84	0.80	0.78
	Retail	*alpha*	−0.75	−0.73	−0.70	−0.70	−0.67	−0.62	−0.56	−0.55	−0.58	−0.40
			(0.21)	(0.24)	(0.26)	(0.26)	(0.25)	(0.26)	(0.26)	(0.26)	(0.22)	(0.26)
		beta	0.97	0.96	0.96	0.95	0.94	0.93	0.92	0.92	0.91	0.89

Panel B: Value-weighted performance of the x% worst performing funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.24	−0.30	−0.32	−0.40	−0.51	−0.51	−0.65	−1.01	−0.94
			(0.64)	(0.64)	(0.64)	(0.64)	(0.60)	(0.56)	(0.58)	(0.46)	(0.24)	(0.43)
		beta	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	Retail	*alpha*	−0.60	−0.63	−0.72	−0.77	−0.80	−0.79	−0.88	−1.04	−1.08	**−1.39**
			(0.27)	(0.27)	(0.26)	(0.27)	(0.27)	(0.27)	(0.27)	(0.19)	(0.17)	(0.04)
		beta	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.98
Fixed Income	Inst.	*alpha*	0.26	0.26	0.13	0.23	0.14	0.13	0.16	0.14	−0.08	−0.26
			(0.77)	(0.78)	(0.96)	(0.85)	(0.96)	(0.96)	(0.94)	(0.96)	(0.96)	(0.80)
		beta	0.88	0.89	0.90	0.91	0.91	0.93	0.94	0.96	0.96	0.98
	Retail	*alpha*	−0.75	−0.77	−0.84	−0.93	−0.96	−0.96	−0.96	−0.84	−0.93	*−0.84*
			(0.17)	(0.17)	(0.17)	(0.17)	(0.15)	(0.15)	(0.14)	(0.17)	(0.11)	(0.09)
		beta	0.97	0.98	0.99	1.01	1.02	1.02	1.02	1.03	1.02	1.00

Table 8: Bivariate sorts on the performance and size of the past year

Annualized alpha after fees (in %) and block-bootstrapped multiple hypothesis adjusted p-value (in brackets) for the value-weighted performance of active mutual funds against the investable benchmark. The portfolios of active mutual funds are double-sorted based on the performance (rows) and the size (columns) of the past year. The nine portfolios arise from the 30th and 70th percentiles within the investment category of each fund. For each panel we distinguish between equity (top) and fixed income (bottom) funds. We rebalance the portfolios every year starting in December 1992 to December 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

		Institutional				Retail		
		Small	Medium	Big		Small	Medium	Big
Equity	Winner	0.71	0.83	−0.05	Winner	0.09	−0.45	−0.28
		(0.76)	(0.54)	(0.96)		(0.89)	(0.60)	(0.74)
	Average	0.65	−0.11	−0.22	Average	−1.15	−0.96	−0.66
		(0.74)	(0.96)	(0.93)		(0.17)	(0.17)	(0.44)
	Loser	−0.76	−0.42	−0.65	Loser	*−1.78*	*−1.53*	−0.99
		(0.69)	(0.93)	(0.70)		(0.07)	(0.05)	(0.30)
Fixed Income	Winner	0.59	−0.25	0.25	Winner	−0.61	−0.49	−0.56
		(0.60)	(0.86)	(0.84)		(0.20)	(0.27)	(0.27)
	Average	−0.31	−0.14	0.37	Average	−1.00	−0.86	−0.93
		(0.86)	(0.92)	(0.80)		(0.13)	(0.21)	(0.27)
	Loser	−0.66	−0.13	0.20	Loser	*−1.25*	−1.05	−0.79
		(0.63)	(0.92)	(0.92)		(0.06)	(0.12)	(0.27)

Table 9: Portfolios filtered by the fee of the past year.

Annualized alpha after fees (in %), the corresponding block-bootstrapped multiple hypothesis adjusted p-value (in brackets), and the beta for the value-weighted performance of active mutual funds against the investable benchmark. Panel A shows the value-weighted performance of portfolios consisting of the x% least expensive active mutual funds of the past year. Panel B shows the performance of the x% most expensive active mutual funds. Every December, the portfolios are rebalanced to exclude a certain percentage of active funds. The data sample ranges from 1993 to 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

Panel A: Value-weighted performance of the x% least expensive funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.13	−0.11	−0.05	−0.04	0.09	0.04	0.17	0.49	0.83
			(0.66)	(0.82)	(0.84)	(0.96)	(0.96)	(0.86)	(0.96)	(0.74)	(0.26)	(0.15)
		beta	1.00	1.00	0.99	1.00	0.99	0.99	1.00	1.00	0.99	0.97
	Retail	*alpha*	−0.60	−0.50	−0.44	−0.40	−0.33	−0.24	−0.15	−0.04	0.03	0.30
			(0.23)	(0.35)	(0.42)	(0.48)	(0.58)	(0.66)	(0.80)	(0.97)	(0.97)	(0.62)
		beta	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.97	0.96
Fixed Income	Inst.	*alpha*	0.26	0.32	0.33	0.33	0.35	0.33	0.35	0.39	0.25	0.01
			(0.69)	(0.61)	(0.61)	(0.56)	(0.52)	(0.52)	(0.46)	(0.37)	(0.61)	(0.96)
			0.88	0.88	0.87	0.87	0.86	0.86	0.86	0.86	0.81	0.72
	Retail	*alpha*	−0.75	−0.71	−0.66	−0.63	−0.61	−0.56	−0.55	−0.53	−0.42	−0.40
			(0.17)	(0.19)	(0.20)	(0.20)	(0.22)	(0.25)	(0.26)	(0.28)	(0.36)	(0.36)
			0.97	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.94	0.94

Panel B: Value-weighted performance of the x% most expensive funds

			All	90%	80%	70%	60%	50%	40%	30%	20%	10%
Equity	Inst.	*alpha*	−0.23	−0.36	−0.52	−0.50	−0.48	−0.78	−0.65	−0.94	−0.88	−1.14
			(0.50)	(0.36)	(0.30)	(0.32)	(0.34)	(0.20)	(0.32)	(0.22)	(0.30)	(0.20)
		beta	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.02	1.03	1.04
	Retail	*alpha*	−0.60	**−0.99**	**−1.28**	**−1.53**	**−1.70**	**−1.82**	**−1.94**	**−2.10**	**−2.48**	**−2.83**
			(0.15)	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
		beta	0.99	1.00	1.01	1.01	1.01	1.02	1.02	1.02	1.02	1.01
Fixed Income	Inst.	*alpha*	0.26	0.31	0.22	0.10	0.10	0.15	0.06	0.08	−0.11	−0.26
			(0.66)	(0.60)	(0.77)	(0.95)	(0.95)	(0.92)	(0.97)	(0.97)	(0.97)	(0.84)
			0.88	0.89	0.89	0.90	0.90	0.92	0.94	0.94	0.94	0.88
	Retail	*alpha*	−0.75	−0.83	−0.91	−0.92	−0.95	−1.02	−1.02	−1.08	−1.37	−1.46
			(0.16)	(0.16)	(0.14)	(0.14)	(0.14)	(0.14)	(0.16)	(0.16)	(0.14)	(0.14)
			0.97	0.97	0.98	0.98	0.99	1.00	1.00	1.00	1.02	1.00

Table 10: Bivariate sorts on the performance and fee of the past year

Annualized alpha after fees (in %) and block-bootstrapped multiple hypothesis adjusted p-value (in brackets) for the value-weighted performance of active mutual funds against the investable benchmark. The portfolios of active mutual funds are double-sorted based on the performance (rows) and the fee (columns) of the past year. The nine portfolios arise from the 30th and 70th percentiles within the investment category of each fund. For each panel we distinguish between equity (top) and fixed income (bottom) funds. We rebalance the portfolios every year starting in December 1992 to December 2016. Alphas with p-values below 10% are in italics and p-values below 5% are in bold.

		Institutional				Retail		
		High Fee	Medium	Low Fee		High Fee	Medium	Low Fee
Equity	Winner	−0.03	−0.19	0.44	Winner	*−1.78*	−1.03	0.14
		(0.99)	(0.99)	(0.92)		(0.08)	(0.16)	(0.93)
	Average	−1.15	−0.16	−0.14	Average	**−2.13**	*−1.24*	−0.18
		(0.33)	(0.99)	(0.99)		(0.00)	(0.06)	(0.93)
	Loser	−1.61	−1.41	0.25	Loser	**−2.38**	*−1.55*	−0.43
		(0.33)	(0.23)	(0.99)		(0.00)	(0.05)	(0.81)
Fixed Income	Winner	0.10	0.33	0.17	Winner	−0.85	−0.60	−0.42
		(1.00)	(0.77)	(0.91)		(0.37)	(0.35)	(0.37)
	Average	0.17	−0.01	0.49	Average	−1.20	−0.96	−0.74
		(1.00)	(1.00)	(0.57)		(0.35)	(0.22)	(0.37)
	Loser	−0.07	0.04	0.40	Loser	−1.17	−0.79	−0.71
		(1.00)	(1.00)	(0.91)		(0.22)	(0.35)	(0.37)

Table 11: Percentage of rejections under the standard and robust alpha tests

Rejections (in %) based on the normal distribution (Standard), Newey–West (NW), HC3, standard resampling (Robust-SR), and the block resampling (Robust-BR) standard errors. For the single mutual fund analysis, we show the percentage of rejections for all 52,526 active mutual funds for the single hypothesis and for the multiple hypothesis analysis based on controlling the FDR. For the portfolios of mutual funds, we compare the percentage of rejections for the 63 portfolios where we adjust for multiple tries based on the FWER. For the standard tests, the FWER method of Romano and Wolf (2016) is not applicable since it requires the bootstrapped t-values.

	Single Mutual Funds				**Portfolios of Mutual Funds**			
Single Hypothesis	Unskilled		Skilled		Unskilled		Skilled	
	<5%	<10%	<5%	<10%	<5%	<10%	<5%	<10%
Standard	11.2	16.2	5.4	8.1	17.5	20.3	7.8	11.1
NW	12.3	17.5	5.8	8.7	19.0	20.3	4.8	7.9
HC3	10.4	15.3	5.0	7.7	17.5	20.3	7.8	11.1
Robust - SR	11.1	16.4	5.3	8.1	19.0	20.3	4.7	7.9
Robust - BR	9.0	13.9	4.3	6.9	15.9	18.7	1.6	6.3
Multiple Hypothesis	FDR				FWER			
	Unskilled		Skilled		Unskilled		Skilled	
	<5%	<10%	<5%	<10%	<5%	<10%	<5%	<10%
Standard	4.1	7.9	1.4	2.2	not applicable			
NW	4.9	9.3	1.3	2.2	not applicable			
HC3	3.2	6.4	1.3	1.8	not applicable			
Robust - SR	3.3	7.0	0.9	1.6	6.4	9.5	0	0
Robust - BR	1.4	3.8	0.3	0.8	4.7	6.4	0	0

Table A.1: Summary statistics for mutual fund investment categories

Average number (Avg Number) of funds, average total net assets in million USD (Avg TNA), average annual net return in USD (Avg Net Ret), average number of years the fund is in the database (Avg Years), and the first appearance of an index fund, for the time period from December 1991 to December 2016, for all available investment categories (Global Category) within the Morningstar database. We only include funds within the "Global Broad Category Group" equity or fixed income that are flagged as "Open-End" or "Exchange-Traded" funds. The average corresponds to the mean of cross-sectional monthly attributes.

in USD		Avg Number		Avg TNA		Avg Net Ret		Avg Years		First Index
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	Fund
Mexico Fixed Income	*Retail*	182.3	2.0	233.7	137.4	-1.89%	1.28%	5.8	7.5	May 09
	Inst.	363.2	25.8	435.8	685.3	6.73%	6.17%	7.1	6.0	Jul 94
Global Equity Large Cap	*Retail*	1,093.9	57.6	466.4	603.6	6.22%	5.99%	7.5	5.6	Jan 92
	Inst.	22.7	6.2	80.0	26.4	8.93%	9.33%	6.8	6.7	Feb 04
Mexico Equity	*Retail*	26.0	10.6	70.7	297.7	7.01%	9.55%	7.0	8.5	Feb 04
	Inst.	329.7	21.6	295.9	241.4	4.36%	3.85%	5.2	5.2	Apr 03
Global Fixed Income	*Retail*	1,104.1	33.9	201.9	242.3	2.63%	2.68%	5.5	4.4	Apr 05
	Inst.	211.9	24.5	126.2	427.5	6.97%	6.34%	6.0	6.0	Jan 98
Europe Equity Large Cap	*Retail*	812.6	73.8	237.7	425.9	6.71%	7.33%	7.9	6.6	Jan 92
	Inst.	349.7	19.1	298.7	638.7	2.43%	3.40%	6.0	6.6	Apr 04
Euro Fixed Income	*Retail*	1,103.6	23.7	354.7	241.2	3.83%	5.71%	8.3	5.6	Jan 92
	Inst.	234.2	13.3	252.6	459.8	11.88%	11.12%	9.6	9.5	Oct 92
US Equity Small Cap	*Retail*	456.0	43.0	280.5	580.4	10.78%	10.85%	11.0	10.2	Jan 92
	Inst.	165.0	2.1	251.2	236.9	3.86%	0.04%	5.5	5.1	Jun 07
Global Equity	*Retail*	382.6	19.5	706.9	224.1	6.68%	5.65%	9.0	5.7	Aug 95
High Yield Fixed Income	*Retail*	954.2	16.7	307.3	197.2	4.57%	3.24%	5.1	3.6	Dec 07
	Inst.	214.3	4.2	237.7	78.5	2.16%	3.09%	4.3	4.1	Apr 05
Other Fixed Income	*Retail*	651.8	10.8	202.0	59.6	2.12%	2.39%	4.6	3.3	Apr 05
	Inst.	188.1	10.8	254.7	848.5	10.89%	11.92%	9.0	7.5	Dec 92
US Equity Mid Cap	*Retail*	487.2	39.3	450.5	778.7	10.02%	11.00%	10.5	7.5	Jan 92
	Inst.	78.7	10.5	95.9	665.1	7.49%	6.10%	6.5	7.0	Dec 98
Other Europe Equity	*Retail*	547.9	68.6	281.7	209.3	9.03%	8.24%	9.1	8.0	Jan 92
	Inst.	22.2	2.4	25.7	46.4	5.59%	3.74%	7.3	10.1	Mar 04
Financials Sector Equity	*Retail*	113.0	32.8	138.8	114.9	5.29%	5.60%	8.7	6.9	Feb 01
	Inst.	73.7	1.6	131.6	11.0	3.69%	3.65%	5.6	6.6	Jan 09
Africa Fixed Income	*Retail*	90.5	1.7	90.7	22.5	5.02%	5.81%	6.3	3.4	Nov 08
Islamic Equity	*Retail*	70.5	9.5	76.4	31.0	3.65%	3.29%	6.8	7.3	Feb 07
	Inst.	69.1	1.6	56.2	24.4	4.08%	4.84%	5.3	4.4	Mar 07
Other Sector Equity	*Retail*	262.4	40.7	124.0	136.6	4.46%	2.02%	6.2	5.1	Nov 05

Table A.1 (continued)

in USD		Avg Number *Active*	Avg Number *Index*	Avg TNA *Active*	Avg TNA *Index*	Avg Net Ret *Active*	Avg Net Ret *Index*	Avg Years *Active*	Avg Years *Index*	First Index Fund
Africa Equity	Inst.	106.9	6.8	70.7	12.3	0.66%	2.53%	5.5	4.4	Nov 07
	Retail	145.6	17.6	94.6	71.4	12.94%	13.52%	7.3	6.5	Apr 03
Technology Sector Equity	Inst.	38.9	5.3	57.3	29.2	8.94%	7.29%	6.5	9.6	Apr 04
	Retail	165.8	20.7	295.4	182.8	12.51%	11.56%	9.3	8.0	Jan 92
Energy Sector Equity	Inst.	40.5	1.9	103.9	145.9	6.84%	9.06%	5.7	7.8	Nov 04
	Retail	120.9	22.2	219.4	249.5	6.83%	7.39%	8.0	6.7	Jul 00
US Equity Large Cap Growth	Inst.	255.3	6.2	378.6	757.8	6.57%	5.32%	8.8	7.2	Jun 98
	Retail	551.4	21.8	1'229.6	809.9	8.65%	9.82%	11.0	8.1	Dec 92
US Equity Large Cap Value	Inst.	194.3	4.8	415.7	863.6	7.10%	7.16%	8.3	8.0	Aug 98
	Retail	381.8	21.8	1'006.4	758.0	8.62%	10.15%	10.6	6.9	Dec 92
US Fixed Income	Inst.	403.3	16.5	573.4	1'387.8	4.92%	5.35%	9.7	8.6	Jan 92
	Retail	843.9	50.5	540.4	1'540.5	4.59%	5.26%	10.6	7.6	Jan 92
Other Europe Fixed Income	Inst.	81.6	13.3	233.5	1'000.8	5.85%	6.63%	6.8	7.7	Nov 01
	Retail	263.2	7.7	327.7	185.2	4.30%	5.32%	8.2	6.9	Mar 98
US Equity Large Cap Blend	Inst.	233.1	54.6	289.9	2'019.1	8.68%	9.27%	8.4	9.8	Jan 92
	Retail	692.4	146.2	629.6	1'912.6	7.84%	8.98%	9.1	8.7	Jan 92
Asia Equity	Inst.	23.4	1.2	86.3	750.2	3.19%	3.91%	6.1	10.2	Jun 00
	Retail	111.5	6.2	142.0	758.8	4.66%	2.69%	9.1	6.7	Jan 92
Real Estate Sector Equity	Inst.	163.0	12.9	221.8	359.3	9.28%	8.21%	7.1	6.0	Feb 04
	Retail	268.8	26.8	162.3	628.0	9.86%	10.51%	8.2	6.3	Jun 96
Inflation Linked	Inst.	108.6	8.2	303.9	222.0	3.07%	3.46%	6.5	4.8	Feb 04
	Retail	160.9	15.8	284.4	391.6	4.03%	4.60%	7.8	6.2	Dec 98
Emerging Markets Fixed Income	Inst.	457.1	2.5	261.1	205.3	-0.82%	4.03%	2.8	1.8	Jun 13
	Retail	457.5	9.2	174.8	500.0	5.40%	7.32%	5.5	3.8	Mar 04
Emerging Markets Equity	Inst.	231.6	10.6	372.5	495.3	8.64%	8.56%	6.3	4.7	Jul 00
	Retail	417.3	29.5	214.4	427.7	7.69%	7.13%	7.5	5.1	May 92
Asia ex-Japan Equity	Inst.	115.1	10.1	155.3	169.1	12.09%	11.28%	6.0	7.3	Apr 03
	Retail	282.9	16.4	176.6	96.0	5.82%	7.09%	7.3	6.3	Nov 94
Greater China Equity	Inst.	94.1	2.3	71.7	43.5	10.84%	11.45%	5.2	4.5	Apr 09
	Retail	234.5	94.5	188.5	1'084.6	10.27%	10.73%	5.7	4.1	Jan 01
Japan Equity	Inst.	85.5	10.8	129.3	253.5	1.78%	2.31%	5.7	5.4	May 00
	Retail	300.9	37.6	262.0	301.1	3.84%	3.28%	6.9	5.9	Feb 98
UK Equity Large Cap	Inst.	59.1	8.1	324.4	542.7	5.32%	4.55%	6.1	6.4	Jan 06
	Retail	170.6	36.8	423.9	445.4	3.86%	3.52%	6.3	7.3	Nov 99

Table A.1 (continued)

in USD		Avg Number		Avg TNA		Avg Net Ret		Avg Years		First Index
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	Fund
Global Equity Mid/Small Cap	Inst.	103.3	3.3	344.8	223.4	11.47%	13.73%	4.9	4.3	May 09
	Retail	288.3	9.8	230.3	223.5	4.34%	6.45%	6.4	5.2	Jul 06
Asia Fixed Income	Inst.	30.6	2.5	85.6	292.8	3.89%	1.81%	4.6	9.9	Apr 05
	Retail	268.6	6.8	163.8	110.8	5.72%	4.50%	4.9	3.7	Jan 06
Cons. Goods & Serv. Sect. Eq.	Inst.	16.5	9.8	85.4	23.7	8.22%	7.30%	5.3	9.7	Mar 04
	Retail	97.0	32.0	98.2	210.0	7.41%	6.28%	7.1	7.4	Jul 00
Sterling Fixed Income	Inst.	51.7	6.1	275.6	98.6	2.63%	2.04%	5.5	7.1	Apr 05
	Retail	169.7	15.3	447.3	457.7	1.72%	2.18%	6.2	5.8	Apr 05
Europe Equity Mid/Small Cap	Inst.	130.1	1.3	103.5	64.1	7.95%	6.91%	2.7	2.7	Dec 12
	Retail	350.2	7.8	133.1	49.7	7.57%	8.75%	7.6	6.8	Jun 01
Latin America Equity	Inst.	31.3	1.0	127.5	4.6	1.41%	-0.42%	4.1	7.6	Aug 07
	Retail	86.1	10.3	140.3	764.1	8.07%	10.33%	8.1	6.3	Aug 00
Natural Resources Sector Equity	Inst.	37.5	2.0	117.1	45.8	7.65%	8.22%	6.8	8.5	Mar 04
	Retail	107.5	14.5	167.2	162.9	7.63%	6.51%	8.9	6.6	Apr 94
Brazil Equity	Retail	51.4	6.2	54.5	21.9	-0.17%	3.29%	6.5	4.9	Aug 07
India Equity	Inst.	33.7	1.0	89.1	26.0	16.92%	17.15%	1.3	2.4	Sep 08
	Retail	169.7	9.4	169.3	233.7	8.36%	7.66%	7.5	5.2	Jan 07
Utilities Sector Equity	Inst.	15.0	1.6	64.8	143.5	8.51%	8.42%	8.6	9.9	May 04
	Retail	32.6	8.1	644.6	308.8	7.38%	8.80%	11.3	8.1	Jan 92
Healthcare Sector Equity	Inst.	32.8	4.5	77.9	62.1	9.76%	6.76%	6.5	9.7	Mar 04
	Retail	173.5	22.2	415.9	342.7	7.25%	7.11%	8.3	6.1	Jun 00
UK Equity Mid/Small Cap	Inst.	32.5	1.0	151.4	61.7	0.89%	-0.62%	1.2	1.9	Feb 15
	Retail	143.2	4.6	328.5	112.0	6.56%	6.65%	6.5	5.7	Jan 06
Communications Sector Equity	Inst.	4.6	4.9	9.5	8.9	7.25%	6.67%	6.0	9.5	Apr 05
	Retail	39.5	13.7	148.8	73.3	5.73%	8.87%	7.9	9.1	Oct 01
Korea Equity	Inst.	83.5	5.3	33.9	29.9	6.93%	6.05%	4.5	4.4	Mar 07
	Retail	280.0	41.0	96.9	57.1	13.67%	13.23%	8.9	5.7	May 01
Asia Pacific Fixed Income	Inst.	21.6	1.1	108.7	22.3	3.05%	3.75%	3.7	5.5	Sep 06
	Retail	120.9	9.6	39.6	130.1	2.60%	3.34%	7.2	5.9	May 05
Thailand Equity	Retail	117.5	8.8	47.0	49.0	17.88%	17.17%	10.9	9.3	Jan 01
Other Asia Equity	Retail	78.6	3.5	89.8	30.7	1.46%	1.06%	5.2	6.3	Jan 08
Precious Metals Sector Equity	Retail	76.8	5.9	196.1	376.0	12.31%	10.76%	9.6	6.1	Jan 92
Canadian Equity Large Cap	Inst.	4.7	2.8	34.0	216.2	5.03%	11.24%	4.9	6.5	Apr 03
	Retail	159.2	13.3	368.8	379.4	8.61%	8.79%	11.1	6.6	Jan 92

Table A.1 (continued)

in USD		Avg Number		Avg TNA		Avg Net Ret		Avg Years		First Index
		Active	*Index*	*Active*	*Index*	*Active*	*Index*	*Active*	*Index*	Fund
Thailand Fixed Income	Retail	82.8	1.0	126.6	165.0	3.60%	6.25%	7.2	10.8	Mar 06
South American Equity	Retail	35.4	1.2	39.0	16.9	4.51%	4.32%	9.3	6.8	Jan 06
	Inst.	14.6	1.8	33.3	367.9	7.26%	5.58%	5.1	6.5	Sep 09
Other Equity	Retail	42.5	25.3	132.9	240.9	8.27%	7.27%	8.8	6.6	Apr 96
	Inst.	6.0	5.7	31.7	16.8	7.59%	6.41%	6.6	9.0	Jan 06
Industrials Sector Equity	Retail	29.2	23.9	123.6	188.6	8.96%	10.50%	8.9	7.6	Oct 01
	Inst.	4.9	1.0	73.5	56.7	14.47%	13.70%	7.8	4.0	Dec 08
Australia & New Zealand Eq.	Retail	10.0	1.9	149.3	42.1	8.03%	8.92%	8.0	3.8	Feb 05
Canada Fixed Income	Retail	88.9	12.9	296.2	1'064.0	4.91%	5.94%	9.5	6.5	Jan 92
	Inst.	3.4	1.0	61.6	13.4	15.31%	14.31%	5.5	6.4	Apr 09
Singapore Equity	Retail	7.6	3.0	110.9	159.1	10.64%	10.25%	7.4	8.9	May 02
Canadian Eq. Mid/Small Cap	Retail	81.1	2.8	179.6	111.6	5.84%	1.42%	7.5	5.4	Apr 07
	Inst.	3.0	1.0	23.6	4.3	37.40%	29.26%	1.7	2.3	Dec 08
Taiwan Equity	Retail	135.2	9.9	55.9	363.3	10.42%	9.17%	11.1	7.4	Jul 03
Australia Fixed Income	Retail	6.7	1.2	195.8	26.5	3.24%	3.96%	2.8	3.8	Jun 10
Malaysia Fixed Income	Retail	52.9	1.0	64.5	205.8	1.81%	0.75%	6.5	9.3	Oct 07

Table C.1: Empirical rejection probabilities: Type I errors

Empirical rejection probabilities for the nominal levels $\alpha = 10\%$, $\alpha = 5\%$, and $\alpha = 1\%$ for the standard (Stand), Newey–West (NW) with a bandwith of $4 \times (T/100)^{2/9}$, HC3, and our bootstrapped (Boot) significance test that evaluates the optimal block size by the method of Politis and White (2004) and the correction of Patton et al. (2009). The data was generated by sampling from the realized returns with a circular bootstrap (Boot-x) and block sizes of $x = \{1, 3, 6\}$. The simulation study includes ten US mutual funds that exhibit the entire return history from 1992 to 2016 in the Morningstar database. We sample 1,000 paths for each fund and DGP and set the alpha under the null hypothesis to the true observed alpha. We show the results for the one-factor "CAPM," three-factor "FF3," and five-factor "FF5" model with the factors "market," "size," and "value" of Fama and French (1992), and also the "momentum" of Jegadeesh and Titman (1993), and "betting against beta" factor of Frazzini and Pedersen (2014). We highlight the p-values closest to the nominal value of the test. Because the null hypothesis is true for all of the simulations, the true rejection probabilities should be equal to the nominal level of the test.

	Nominal	CAPM				FF3				FF5			
DGP	Level	*Stand*	*NW*	*HC3*	*Boot*	*Stand*	*NW*	*HC3*	*Boot*	*Stand*	*NW*	*HC3*	*Boot*
Boot-1	$\alpha = 0.10$	0.112	0.109	**0.099**	0.102	0.114	0.113	**0.099**	0.105	0.120	0.117	**0.096**	0.106
	$\alpha = 0.05$	0.061	0.059	**0.049**	0.053	0.062	0.060	**0.050**	0.056	0.064	0.061	0.046	**0.053**
	$\alpha = 0.01$	0.012	0.012	0.008	**0.010**	0.016	0.016	**0.012**	0.013	0.016	0.017	**0.011**	0.013
Boot-3	$\alpha = 0.10$	0.142	0.119	0.125	**0.111**	0.137	0.123	0.120	**0.106**	0.135	0.118	0.111	**0.105**
	$\alpha = 0.05$	0.084	0.065	0.070	**0.058**	0.078	0.068	0.066	**0.060**	0.076	0.062	0.058	**0.054**
	$\alpha = 0.01$	0.024	0.017	0.017	**0.013**	0.025	0.020	0.019	**0.015**	0.023	0.018	0.017	**0.013**
Boot-6	$\alpha = 0.10$	0.158	0.126	0.141	**0.115**	0.140	0.124	0.124	**0.112**	0.148	0.126	0.123	**0.114**
	$\alpha = 0.05$	0.098	0.071	0.085	**0.063**	0.078	0.071	0.066	**0.062**	0.088	0.071	0.068	**0.062**
	$\alpha = 0.01$	0.036	0.021	0.029	**0.017**	0.025	0.021	0.020	**0.018**	0.027	0.022	0.020	**0.015**

Table C.2: Empirical rejection probabilities: Family wise error rates (FWER)

Empirical rejection probabilities for the nominal levels $\alpha = 10\%$, $\alpha = 5\%$, and $\alpha = 1\%$ and the multiple hypothesis framework of Romano and Wolf (2005a,b, 2016) controlling the FWER based on the bootstrapped (Boot-x) significance test with block sizes of $x = \{1, 3, 6, 9, 12\}$. The DGP is a circular bootstrap (Boot-x) with an optimal block size of $x = \{1, 3, 6\}$. The simulation study includes the 17 retail investment categories with a history from 1993 to 2016 from Section 2 with the investable one-factor benchmark model that is based on the value-weighted return of index funds. For each portfolio and DGP we sample 1,000 paths and set the alpha under the null hypothesis to the true observed alpha. We highlight the p-values closest to the nominal value of the test. Because for all the simulations the null hypothesis is true, the true rejection probabilities should be equal to the nominal level of the test.

DGP	Nominal Level	Boot-1	Boot-3	Boot-6	Boot-9	Boot-12
Boot-1	$\alpha = 0.10$	0.132	**0.119**	0.096	0.071	0.050
	$\alpha = 0.05$	0.066	**0.052**	0.039	0.026	0.020
	$\alpha = 0.01$	0.015	**0.009**	0.006	0.004	0.002
Boot-3	$\alpha = 0.10$	0.145	0.124	**0.098**	0.082	0.059
	$\alpha = 0.05$	0.081	0.062	**0.048**	0.040	0.030
	$\alpha = 0.01$	0.028	0.016	**0.010**	0.004	0.002
Boot-6	$\alpha = 0.10$	0.132	0.114	**0.087**	0.066	0.046
	$\alpha = 0.05$	0.073	**0.051**	0.036	0.022	0.012
	$\alpha = 0.01$	0.020	**0.008**	0.004	0.000	0.001