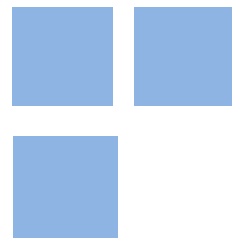


Unskilled Fund Managers: Replicating Active Fund Performance With Few ETFs

Elias Cavalcante Júnior

Rodrigo De-Losso

Fernando Moraes



Unskilled Fund Managers: Replicating Active Fund Performance With Few ETFs

Elias Cavalcante Júnior (e.cavalcante@usp.br)

Fernando Moraes (fernandotm@al.insper.edu.br)

Rodrigo De-Losso (delosso@usp.br)

Research Group: NEFIN

Abstract

This paper use Exchange Traded Funds (ETFs) instead of risk factors as benchmarks to examine active mutual fund performance distribution. While transaction costs are included in the ETF returns, that is not true regarding risk factors, making it more challenging to characterize extraordinary performances via alphas. Assessments are based on the proportion of skilled funds, defined as positive-alpha funds. Such a proportion is calculated taking into account potential false discoveries and employing the method devised by Barras et al. (2010). After evaluating several ETF combinations, we conclude that sets of 3 to 5 ETFs replicate most levels of active fund performance. Finally, we propose specific ETF selection algorithms, whereby we estimate that 95% of active management funds fail to generate value for their investors. Alphas calculated with ETFs are higher than those using risk factors, but the difference is similar to the transaction costs required for investing in risk factor portfolios (Frazzini et al., 2012).

Keywords: Mutual funds, performance measures, ETF, risk factors.

JEL Codes: G11, G2, G23.

Unskilled fund managers: replicating active fund performance with few ETFs

Elias Cavalcante-Filho¹

Fernando Tassinari Moraes²

Rodrigo De-Losso³

June 25, 2020

Abstract

This paper use Exchange Traded Funds (ETFs) instead of risk factors as benchmarks to examine active mutual fund performance distribution. While transaction costs are included in the ETF returns, that is not true regarding risk factors, making it more challenging to characterize extraordinary performances via alphas. Assessments are based on the proportion of skilled funds, defined as positive-alpha funds. Such a proportion is calculated taking into account potential false discoveries and employing the method devised by Barras *et al.* (2010). After evaluating several ETF combinations, we conclude that sets of 3 to 5 ETFs replicate most levels of active fund performance. Finally, we propose specific ETF selection algorithms, whereby we estimate that 95% of active management funds fail to generate value for their investors. Alphas calculated with ETFs are higher than those using risk factors, and the difference is similar to the transaction costs required for investing in risk factor portfolios (Frazzini *et al.*, 2012).

Keywords: Mutual funds, performance measures, ETF, risk factors.

JEL Codes: G11, G2, G23.

¹ E-mail: e.cavalcante@usp.br.

² E-mail: fernandotm@al.insper.edu.br.

³ E-mail: delosso@usp.br.

1. Introduction

In this paper, we use Exchange Traded Funds (ETFs) instead of risk factors as benchmarks to examine active mutual fund performance distribution. While transaction costs are included in the ETF returns, that is not true regarding risk factors, making it more challenging to characterize extraordinary performances via alphas. We base our assessment on the False Discovery Rate (FDR) method developed by Barras *et al.* (2010), employing a sample of active U.S. asset mutual funds from 2005 to 2019.

In regards to alphas, our results are distinct from those obtained using risk factors, and the difference is nearly the same as the transaction costs required for investing in risk factor portfolios (Frazzini *et al.*, 2012). On the other hand, akin to the results obtained by approaches that use risk factors (Barras *et al.*, 2010), we estimate that approximately 95% of the industry fails to generate value for its investors. Moreover, we point to the fact that a random selection of ETFs is unsuited to outperform active funds; however, this can be achieved by applying specific guided methods to select a set of ETFs.

Using passive funds as benchmarks for active funds is not new. Malkiel (1995) argues that active managers should use passive funds as opportunity cost. Sharpe (1992) seminal paper explain fund returns by twelve different asset class index. The author shows evidence that a constrained regression produces nearly the same results as a regular regression, which leads to the conclusion that individual investors are able to replicate certain funds' performance with a limited-leverage asset class index portfolio.

Although ETFs have shown massive growth in terms of capital allocation, only a small number of studies⁴ have taken them into account. Using the Bloomberg database, we are able to find over 2,000 ETFs negotiated in the U.S. stock exchanges, which covers a significant variety of asset classes, geographic locations and investment strategies. By looking at ETFs focused only in the U.S. equity domestic markets, our dataset displays 875 funds, accounting for an overall sum of 2.1 trillion dollars of assets under management – AUM – as of December, 2019. The Lettau & Madhavan's (2018) survey suggests that this substantial growth might stem from the fact that ETFs are liquid instruments that include transaction costs, in addition to being able to replicate a wide range of investment strategies, including benchmarks.

⁴ See Alexander & Barbosa (2008) and Poterba & Shoven (2002) as examples of ETF-related economic studies.

In order to estimate the proportion of skilled active funds, we choose to focus on individual investors' net-of-fee performance and cost. We single out calculated alpha from net-of-fee excess returns, customarily used⁵ to measure individual investors' economic returns. Therefore, in our terminology, a skilled fund is defined as any fund capable of delivering idiosyncratic returns after it is controlled by certain risks (a set of risk factors or, in our case, selected ETF excess returns).

From an econometric perspective, assessing the proportion of skilled active funds is something far from trivial since we have to deal with lucky and unlucky funds (false positives / Type I error). However, a wide range of papers has already addressed that issue. Fama & French (2010) propose a bootstrap procedure to calculate zero-alpha fund distribution and come to a conclusion on how skill funds are distributed after comparing them to empirical distribution⁶. Pástor & Stambaugh (2002), on the other hand, apply Bayesian techniques, setting priors for unskilled, zero-alpha and skilled fund distributions⁷. Chen *et al.* (2017) assume a parametric mixture model of distribution for alpha funds. We follow the False Discovery Rate (FDR) approach devised by Barras *et al.* (2010). The methodology controls luck funds by calculating their actual proportion within the overall sample. Additionally, it poses the benefits of not requiring prior assumptions about alpha fund distributions and has been widely applied in fund industry research⁸.

We propose specific algorithms that enable investors to choose a parsimonious set of ETFs suited to evaluate actively managed funds. We start with a random ETF selection. Repeating it several times, we measure the average outperforming capacity of a random ETFs group within mutual funds. Second, we propose an algorithm that chooses ETFs based on their ability to outperform mutual funds and to explain fund return variability. At last, applying lasso and forward selection, we select ETFs by looking into the risk factor space.

⁵ See Carhart (1997); Fama & French (2010); Malkiel (1995), Carhart (1997), Barras *et al.* (2010); Crane & Crotty (2018).

⁶ See also Kosowski *et al.* (2006) and Ferson & Chen (2015).

⁷ See also Baks *et al.* (2001) and Avramov & Wermers (2006).

⁸ See Patton & Ramadorai (2013) and Criton & Scaillet (2014) for other empirical studies that apply the FDR estimator in the mutual fund industry.

This paper aims, above all, to help introduce a chapter on ETFs to the discussion of active versus passive fund performance evaluation. Our findings support the current lack of skills existing in the active mutual fund industry.

2. Methodology

To evaluate whether actively managed funds generate value for their investors by outperforming passive investments, we follow standard procedures set forth in academic literature, as well as benchmarked funds against the investment opportunity set that passive investors have to deal with, in this case, the ETF net return.

Following the traditional viewpoint conveyed in the literature (Ardia & Boudt, 2018; Barras *et al.*, 2010; Fama & French, 2010), we define performance as fund managers' ability to generate higher alphas. Therefore, we begin our analysis with the following linear risk factor model:

$$r_{i,t} = \alpha_i + \beta_i \mathbf{ETF}_t + \varepsilon_{i,t} \quad (1)$$

where $r_{i,t}$ is the i mutual fund excess return for period t , \mathbf{ETF}_t is a $(P \times 1)$ vector of P selected ETF fund excess returns, $i \in (1, \dots, I)$, and $t \in (1, \dots, T)$. It is essential to point out that both mutual and ETF fund returns are net-of-fees and expenses measuring individual investors' direct returns.

Based on the model (1), we assume that the actively managed mutual fund population is comprised of three distinct performance categories: unskilled funds ($\alpha < 0$), zero-alpha funds ($\alpha = 0$), and skilled funds ($\alpha > 0$). Therefore, if mutual fund i is a skilled fund ($\alpha_i > 0$), it yields an idiosyncratic positive excess return and outperforms those selected benchmarks.

For any given ETF selection, our main interest is to evaluate the true proportion of skilled mutual funds within our sample. By estimating (1) for $i = (1, \dots, I)$, as is already widely known in the literature, the direct inference of the proportion of skilled funds collecting p-values from $H_i^0: \alpha_i = 0$ against $H_i^1: \alpha_i \neq 0$ turns into an issue since it fails to control false positives (Type I error)⁹. These false positives became known as

⁹ See Barras *et al.* (2010); Chen *et al.* (2017); Fama & French (2010); Kosowski, Timmermann, Wermers, & White (2006); and Pástor & Stambaugh (2002).

lucky and unlucky funds. To address this problem, we apply the FDR (False discovery rate) approach devised by Storey (2002), initially used in the mutual industry fund by Barras *et al.* (2010).

As Andrikogiannopoulou & Papakonstantinou (2019) stress, in order to achieve a meaningful FDR estimator of skilled proportion funds, we need to make sure that individual tests estimated with the model (1) do not yield low power. Therefore, we must choose a sparse set of ETFs to calculate (1) by OLS, including an adequate time series length. As we describe in the following subsections, we employ four different types of ETF selections to achieve a sparse set of predictors.

To compare results attained from our different ETF selections, we set the risk factor model below as our benchmark:

$$r_{i,t} = \alpha_i + \beta_i \mathbf{f}_t + \varepsilon_{i,t} \quad (2)$$

where \mathbf{f}_t is a $(P_f \times 1)$ risk factor vector.

2.1 FDR Approach to Estimate Skilled Proportion Funds

The FDR methodology is used to control false-positive results, estimating their proportion from the overall sample. It sets the proportion of false-positive results classified as skilled (unskilled) funds to F_γ^+ (F_γ^-)¹⁰. By defining π_0 as the proportion of zero-alpha funds for a given γ level of significance, the expected value of F_γ^+ can be written as:

$$E(F_\gamma^+) = \pi_0 \frac{\gamma}{2} \quad (3)$$

After p-values from $H_i^0: \alpha_i = 0$ against $H_i^1: \alpha_i \neq 0$ for the model (1)¹¹ are computed, we proceed to measure the proportion of funds that yield a significant positive-alpha as follows:

$$\hat{S}_\gamma^+ = \frac{\sum_{i \in I_{\hat{\alpha}^+}} \mathbb{I}(\hat{p}_i > \frac{\gamma}{2})}{I} \quad (4)$$

¹⁰ F_γ^+ (F_γ^-) are known in the mutual fund industry literature as lucky (unlucky) funds.

¹¹ See appendix A.1. for a complete description of the p-value estimation procedure.

where $\mathbb{I}(\hat{p}_i > \frac{\gamma}{2})$ is an indicator function equal to 1 if $\hat{p}_i > \frac{\gamma}{2}$ and zero otherwise; and $I_{\hat{\alpha}^+} = \{i: \hat{\alpha}_i > 0\}$. Therefore, the expected proportion of skilled funds in the population, corrected by the proportion of false-positive, can be estimated as:

$$E(\pi_+) = E(S_{\gamma^+}^+) - E(F_{\gamma^+}^+) \quad (5)$$

Hence, with an estimator of the proportion of zero-alpha funds ($\hat{\pi}_0$), and a γ^* optimal significance level, we can calculate $E(\pi_+)$ as follows:

$$\hat{\pi}_+ = \hat{S}_{\gamma^*}^+ - \hat{\pi}_0 \frac{\gamma^*}{2} \quad (6)$$

Since the sum of the proportion of unskilled, zero-alpha and skilled funds must be 1, we can calculate $E(\pi_-)$ with:

$$\hat{\pi}_- = 1 - \hat{\pi}_0 - \hat{\pi}_+ \quad (7)$$

It is worth noting that the entire process from (3) to (7) can be adjusted by estimating $\hat{\pi}_-$ by (6), and $\hat{\pi}_+$ by (7). Therefore, one needs simply to reverse the variable subscriptions +/- . In fact, this decision depends on the number of the elements of sets $I_{\hat{\alpha}^+}$ and $I_{\hat{\alpha}^-}$ ¹². If $I_{\hat{\alpha}^-}$ has more elements than $I_{\hat{\alpha}^+}$, we calculate $\hat{\pi}_-$ using (6), and $\hat{\pi}_+$ using (7), otherwise, we follow the exact same process given by models (3) to (7). In both cases, the optimal significance level (γ^*) is determined by the bootstrap procedure detailed in appendix A.3. The π_0 estimation procedure is described in the following section.

2.1.1 Calculation Procedure for π_0

By definition, and pursuant to null hypothesis $H_i^0: \alpha_i = 0$, the multiple test for $H_i^0: \alpha_i = 0$ against $H_i^1: \alpha_i \neq 0$ generates a series of p-values ($\{p_i\}_{i=1}^I$) that results in a uniform distribution across the $[0,1]$ interval¹³. In other words, if the null hypothesis is true, we can, therefore, state that:

$$H_{0,i}: \alpha_i = 0 \Rightarrow p_i \sim U(0,1) \quad (8)$$

¹² $I_{\hat{\alpha}^-} = \{i: \hat{\alpha}_i < 0\}$.

¹³ Hung, O'Neill, Bauer, & Kohne (1997), and Murdoch, Tsai, & Adcock (2008) submit a feature on the condition of the null hypothesis and p-value distribution.

In our second assumption, we presume that p-values larger than a threshold λ^* can only originate from zero-alpha funds. Hence, considering the p-value empirical distribution, the area given by $\frac{\sum_{i=1}^I \mathbb{I}(\hat{p}_i > \lambda^*)}{I}$ becomes the proportion of the uniform distribution of p-values from zero-alpha funds in the $[\lambda^*, 1]$ interval ¹⁴. In assuming that the area is equal to the $[0,1]$ range, we can determine the proportion of the entire uniform distribution relating to the full p-value mass, thereby computing the π_0 estimator as:

$$\hat{\pi}_0(\lambda^*) = \frac{\sum_{i=1}^I \mathbb{I}(\hat{p}_i > \lambda^*)}{I \times (1 - \lambda^*)} \quad (9)$$

where the optimal threshold level (λ^*) is determined by the bootstrap procedure described in appendix A.2.

2.1.2 Simulated Example

In this section we illustrate the calculation of $\boldsymbol{\pi} = (\pi_-, \pi_0, \pi_+)$ using the FDR approach for a simulated dataset. We were able to generate 1,000 alpha ETFs from $\alpha \sim N(-0.025, 0.01)$ (unskilled funds), 8,000 alphas from $\alpha \sim N(0, 0.01)$ (zero-alpha funds), and 1,000 alphas from $\alpha \sim N(0.025, 0.01)$ (skilled funds). By construction, it is a known fact that the true parameters are $\boldsymbol{\pi} = (0.10, 0.80, 0.10)$.

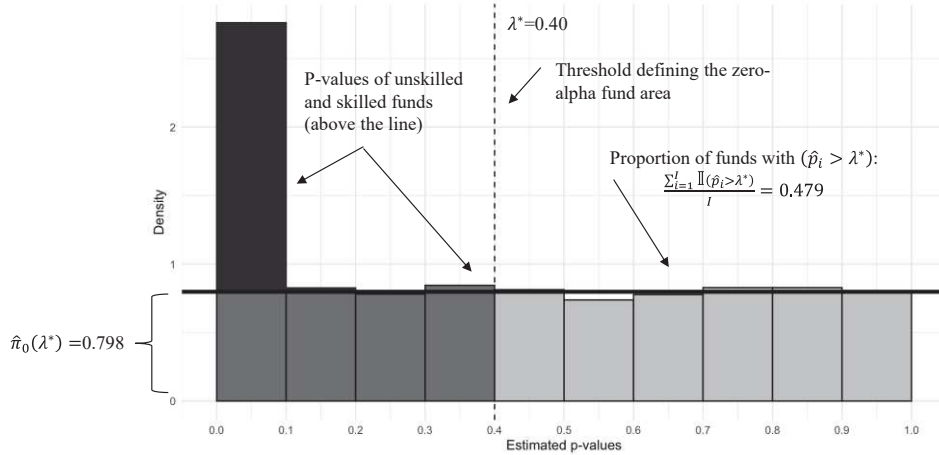
For each generated alpha, we test $H_i^0: \alpha_i = 0$ against $H_i^1: \alpha_i \neq 0$, after which point we compute the estimated p-values (\hat{p}_i), as described in the previous section. Figure 1 plots the histogram of \hat{p}_i , and we realize that the optimal λ^* for this synthetic data is 0.40. The light grey area is the proportion of funds by which $\hat{p}_i > \lambda^*$, and by assumption, we infer that it comes only from zero-alpha funds. Looking at equation (9), we can see that the dark grey area in Figure 1 depicts the deduction of the uniform distribution of zero-alpha funds for the $[0, \lambda^*)$ interval, and since the light grey area is by assumption the uniform distribution area for $[\lambda^*, 1]$, $\hat{\pi}_0(\lambda^*)$ is the result of the sum of the light and dark grey areas. The black area in Figure 1 refers to proposition funds that are either unskilled or skilled.

Panel B of Figure 2 displays t-statistics-related empirical distribution. As one can see, if we fail to correct the proportion of positive significant alphas (\hat{S}_v^+) with the false

¹⁴ I is an indicator function that is equal to 1 if $\hat{p}_i > \lambda^*$ or, otherwise, to zero.

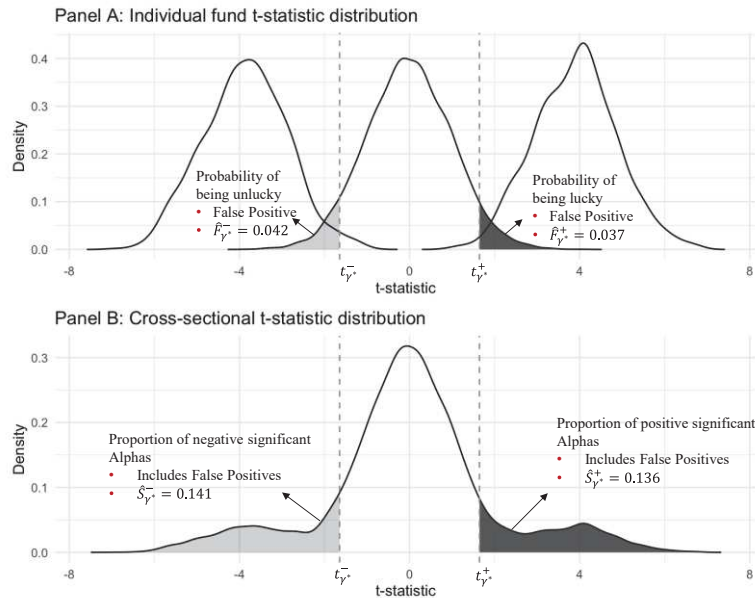
positive ($F_{\gamma^*}^+$ area in Panel A from Figure 2), we would attain a 13.6% proportion of skilled funds instead of a 9.62% proportion for our final calculation of $\hat{\pi}_+$. Table 1 summarizes our procedural results for the simulated dataset, and we can consider $\hat{\pi} = (0.106, 0.789, 0.962)$ to be close to the original parameters of $\pi = (0.10, 0.80, 0.10)$.

Figure 1: Estimated p-values from BY $H_i^0: \alpha_i = 0$ for the simulated dataset



Note: The figure represents the p-value histogram of 10,000 simulated funds. We managed to generate 1,000 alphas from $\alpha \sim N(-0.025, 0.01)$ (unskilled funds), 8,000 alphas from $\alpha \sim N(0, 0.01)$ (zero-alpha funds), and 1,000 alphas from $\alpha \sim N(0.025, 0.01)$ (skilled funds). Original parameters are $\pi = (0.10, 0.80, 0.10)$.

Figure 2: T-statistics distribution from $H_i^0: \alpha_i = 0$ for the simulated dataset



Note: Panel A displays t-statistic distribution in funds across three skill groups (zero-alpha, unskilled, and skilled funds). True non-zero alphas are set to -2.5% , and $+2.5\%$ for unskilled and skilled funds. Panel B depicts the cross-sectional t-statistic distribution. It is a combination of the three distributions in Panel A, where the weight on each distribution depends on the proportion of zero-alpha, unskilled, and skilled funds, whose population values are $\pi_- = 0.10, \pi_0 = 0.80, \pi_+ = 0.10$.

Table 1: FDR approach results for skilled proportion funds in the simulated dataset

Value	π_-	π_0	π_+
Simulated Example			
Proportion	9.62%	79.77%	10.61%
Number	962	7,977	1,061
Std.	0.0036	0.0083	0.0073

Note: The table reports proportions of zero-alpha, skilled, and unskilled (π_0, π_+, π_-) funds based on the risk factor model used for the simulated example set. $\gamma^* = 0.40$ and $\lambda^* = 0.10$ values are obtained with the selection algorithms proposed by (Barras et al., 2010).

2.2 Random ETF Selection

Here, we use a random ETF selection as our first methodology to define the set for the \mathbf{ETF}_t vector for the model (1). Random selection is the most straightforward technique applied to establish sets of ETFs. Furthermore, it provides an average measure of the outperforming capacity of a random group of selected ETFs within mutual funds, but the procedure must be repeated several times. Since \mathbf{ETF}_t is a $(P \times 1)$ vector, we define a dimension threshold of $P^* = 10$ so as $P = 1, \dots, P^*$. By establishing a low level for P^* , we are manually able to set a low dimension for \mathbf{ETF}_t .

For each P dimension, this procedure starts by randomly selecting a b subsample ($\mathbf{ETF}_{(P),(b),t}$) from the original ETF dataset. With this subsample data, we fit the following model to $i = 1, \dots, I$ mutual funds:

$$r_{i,t} = \alpha_i + \beta_i \mathbf{ETF}_{(P),(b),t} + \varepsilon_{i,t} \quad (10)$$

After estimating the model (10), we can compute: $R_{(P),(b),i}^2$, $\hat{\alpha}_{(p),(b),i}$ both for $i = 1, \dots, I$, and $\hat{\boldsymbol{\pi}}_{(P),(b)}$.

We repeat this procedure $B = 500$ times so as $b = 1, \dots, B$ in order to measure the following statistics:

$$R_{(P)}^2 = \frac{1}{BI} \sum_{b=1}^B \sum_{i=1}^I R_{(P),(b),i}^2 \quad (11)$$

$$\hat{\alpha}_{(P)} = \frac{1}{BI} \sum_{b=1}^B \sum_{i=1}^I \hat{\alpha}_{(p),(b),i} \quad (12)$$

$$\hat{\pi}_{(P)} = \frac{1}{B} \sum_{b=1}^B \hat{\pi}_{(P),(b)} \quad (13)$$

Hence, for a selected P of the \mathbf{ETF}_t vector dimension, we calculate the average adjusted- R^2 (11), the average calculated alpha (12), and the average fund category proportion (13) across all subsamples and all mutual funds.

2.3 Proposed ETF Algorithm Selection (PS)

In this section we describe a proposed algorithm selection that chooses ETFs based on their ability to outperform mutual funds (high $\hat{\pi}_+$), as well as their ability to explain fund return variability (high average adjusted- R^2)¹⁵.

We define E as the overall number of ETFs available in our original dataset so that $\mathbf{ETF}_t = (ETF_{(1),t}, \dots, ETF_{(E),t})$. We kick off the first step of our algorithm by calculating:

$$r_{i,t} = \alpha_i + \beta_i ETF_{(e),t} + \varepsilon_{i,t} \quad (14)$$

where $e = (1, \dots, E)$.

For each $ETF_{(e),t}$, we compute the average adjusted- R^2 ($R_{(e)}^2$)¹⁶ and the estimated skilled fund proportion ($\hat{\pi}_{(e),+}$). Additionally, we select a set of ETFs featuring an $R_{(e)}^2$ higher than a $R^{2(*)}$ threshold¹⁷ so that:

$$\mathbf{ETF}_{1,t}^{R^2} = \{e: R_{(e)}^2 \geq R^{2(*)}\} \quad (15)$$

Next, based on the $\mathbf{ETF}_t^{R^2}$ set, we choose the ETF yielding the lowest $\hat{\pi}_{(e),+}$ so that:

$$ETF_{(e_1),t} = \left\{ e: e = \underset{e \in \mathbf{ETF}_{1,t}^{R^2}}{\operatorname{argmin}} \hat{\pi}_{(e),+} \right\} \quad (16)$$

¹⁵ The highest possible adjusted- R^2 is required to address the issues as pointed out by Andrikogiannopoulou & Papakonstantinou (2019).

¹⁶ $R_{(e)}^2 = \frac{\sum_{i=1}^I R_{(e),i}^2}{I}$ where, $R_{(e),i}^2$ is the adjusted R^2 from the fitted model $r_{i,t} = \alpha_i + \beta_i ETF_{(e),t} + \varepsilon_{i,t}$.

¹⁷ We set $R^{2(*)} = 0.85$ for our empirical research.

where $ETF_{(e_1),t}$ is the ETF selected in the first step.

In the second step, we measure:

$$r_{i,t} = \alpha_i + \beta_{i,e_1}ETF_{(e_1),t} + \beta_iETF_{(e),t} + \varepsilon_{i,t} \quad (17)$$

where $e = (1, \dots, E) \setminus \{e_1\}$.

We then select a pair of ETF sets that produce a higher adjusted R^2 than that of the previous sole $ETF_{(e_1),t}$ so that:

$$ETF_{2,t}^{R^2} = \{e: R_{(e)}^2 \geq R_{(e_1)}^2\} \quad (18)$$

Next, we pick the second ETF using:

$$ETF_{(e_2),t} = \left\{ e: e = \underset{e \in ETF_{2,t}^{R^2}}{\operatorname{argmin}} \hat{\pi}_{(e),+} \wedge \hat{\pi}_{(e),+} < \hat{\pi}_{(e_1),+} \right\} \quad (19)$$

Model (16) enables us to conclude that, in order to be selected, the second ETF must also enhance its outperformance across mutual funds by attaining a lower estimated skilled fund proportion.

We proceed with the algorithm until $ETF_{s,t}^{R^2} = \emptyset$, or $ETF_{(e_s),t} = \emptyset$, where s is the number of steps. Therefore, the final selection can be expressed as $ETF_{A,t} = \bigcup_{s=1}^S \{ETF_{(e_s),t}\}$.

2.4 Lasso Selection regarding Risk Factors

Given the well-known fact that risk factors outperform industry mutual funds¹⁸, one can attain a more sophisticated ETF selection by looking into the risk factor space. Based on the Fama & French (2010) and Barras *et al.* (2010), we focus on the four-risk factor model proposed by Carhart (1997), setting our risk factor to $\mathbf{f}_t = (f_{Mkt,t}, f_{SMB,t}, f_{HML,t}, f_{WML,t})$. As highlighted in the Fama (1998), desirable properties from a multifactor risk model like (2) are neither an idiosyncratic risk ($\alpha_i = 0$) nor a satisfactory explanatory power (significant betas and high adjusted- R^2). Consequently, for each f risk factor¹⁹, we look for ETFs that might spam the risk factor space, applying the following model:

¹⁸ See Fama & French (2010) and Barras *et al.* (2010).

¹⁹ $f = (Mkt, SMB, HML, WML)$.

$$f_{f,t} = \beta_f \mathbf{ETF}_t + \varepsilon_{f,t} \quad (20)$$

Since original ETF datasets pose a great number of predictors, as well as the fact that we are interested in selecting a sparse number of predictors, we apply the Lasso estimator²⁰ to (20) with the loss function, as follows:

$$L(\beta_f) = \frac{1}{2} \sum_{t=1}^T (f_{f,t} - \beta_f \mathbf{ETF}_t)^2 + \lambda_f \|\beta_f\|_1 \quad (21)$$

$$\hat{\beta}_{Lasso,f} = \operatorname{argmin}_{\beta_f} L(\beta_f) \quad (22)$$

ETFs selected with risk factor f are defined as $\mathbf{ETF}_{Lasso,f,t} = \{\mathbf{ETF}_t : \hat{\beta}_{Lasso,f} \neq \mathbf{0}\}$. Hence, the final set of ETFs selected by the Lasso procedure can be attained calculating: $\mathbf{ETF}_{Lasso,t} = \cup_{f \in F} \{\mathbf{ETF}_{Lasso,f,t}\}$.

2.5 Forward Selection regarding Risk Factors

We also apply the forward fund selection algorithm as an alternative method to select ETFs in model (20). We define a grid of $k = 1, \dots, K$ ²¹ with K being the threshold for the \mathbf{ETF}_t vector dimension in equation (14). Therefore, the sparsity level in the forward fund selection algorithm is directly controlled by the value of K .

For each risk factor f we start the procedure by setting $k = 1$, and adjusting all possible models to (14). It is worth noting that since $k = 1$, in this first step the number of regressions is exactly the same as the number of elements in the original ETF dataset. We choose the ETF that yields the highest adjusted R^2 . Then, we move on to the second step by setting $k = 2$, and adjusting all possible models to (14) with the ETF selected in the previous step defined as the first predictor, whereas all other remaining ETFs from the original dataset remain potential candidates for the second predictor. Again, we select the ETF set yielding the highest adjusted- R^2 and repeat those same steps until $k = K$, at which point we pick the final K ETF ($\{\mathbf{ETF}_{F,f,t}\}$). As in the Lasso procedure, the final selection is a combination of every selection we make for each risk factor ($\mathbf{ETF}_{F,t} = \cup_{f \in F} \{\mathbf{ETF}_{F,f,t}\}$).

²⁰ See Hastie, Tibshirani, & Friedman (2009) for an extensive description of Lasso estimator properties.

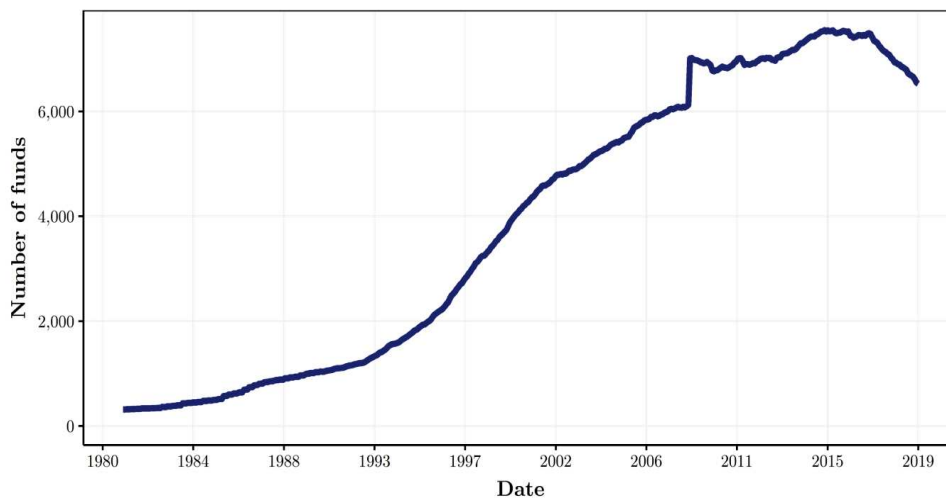
²¹ We select two ETF sets with this methodology, one with $K = 1$, and other with $K = 2$.

3 Data

Fund data is collected from the CRSP Survivor-Bias-Free U.S. Mutual Fund Database²² from January 1980 to December 2019 on a monthly basis. We pick only U.S. domestic equity market funds. To avoid incubation bias, we remove all funds' observations before said funds' AUM reaches 5.0 million 2000 dollars²³. Once the funds are included in the dataset, we keep analyzing them until its AUM drops below 1.0 thousand 2000 dollars. Next, we remove funds that have less than 30 months of observable data. After all adjustments are made, our dataset displays an unbalanced panel with 13,922 funds from January 2005 to December 2019. Risk factor and risk-free rate historical returns were obtained from the French Library²⁴.

Figure 3 depicts the number of analyzed funds in our final dataset universe per month. The figure clearly shows a positive trend in the number of funds over the years. Consequently, most funds are observed in recent years and mainly within our time-sample from January 2005 to December 2019.

Figure 3: Monthly number of funds from January 1980 to December 2019



Note: The figure shows the historical evolution of the 13,922 mutual funds selected from 1980 to 2019. Numbers were gathered from the CRSP U.S. domestic mutual fund database with the AUM ranging from 5.0 million to 1.0 thousand 2000 dollars, including over 30 months of observable data.

Figure 4 displays the evolution in number of ETFs from 1994 to 2019. We can observe that there are more than 800 ETFs available in this period. Moreover, even though

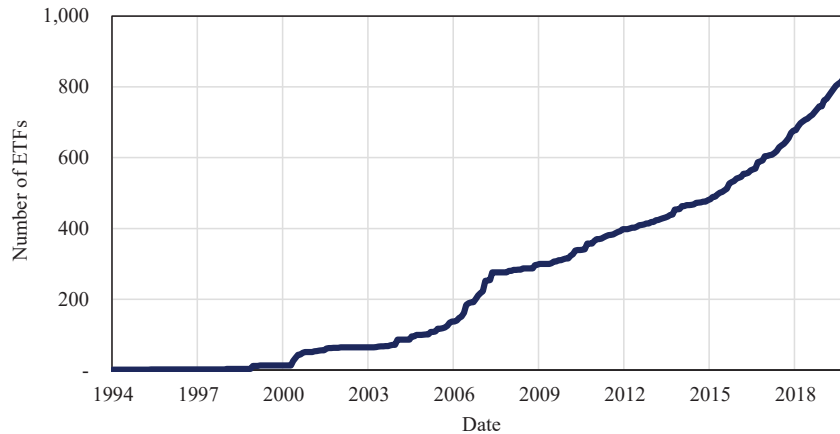
²² See CRSP (2012) for a complete description of this dataset.

²³ Following Berk & van Binsbergen (2015); Fama & French (2010); Harvey & Liu (2019); and Franzoni & Schmalz (2017).

²⁴ https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

ETFs are a considerable market nowadays, they are also a recent financial innovation. Hence, there is a trade-off between the time series length and the number of ETFs available. Under this context, we begin our time-sample in January 2005, which leads to a balanced panel of 100 ETFs issued up to this date.

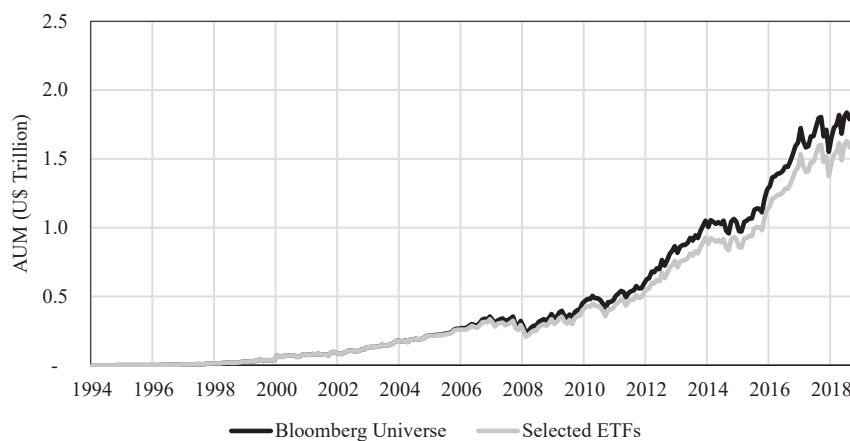
Figure 4: Number of ETF funds in the Bloomberg dataset



Note: The figure displays the number of ETFs available in the Bloomberg database.

In Figure 5 we compare the AUM of the Bloomberg ETF population to our selected sample. The selected ETFs represent more than 85% of the overall AUM database, amounting to over 2.0 trillion dollars of AUM. Here, we must point out that our ETF sample encompasses a significant share of the ETF market, as well as the fact that liquidity is not an issue for individual investors.

Figure 5: ETF AUM from January 1994 to December 2019



Note: The figure displays the evolution of AUM (US\$ trillions) for the Bloomberg ETF sample and the selected ETF.

Table 2 brings statistics for our final dataset universe, ranging from January 2005 to December 2019. Panel A and Panel B consolidate domestic equity mutual funds, as

well as the selected ETF sample, respectively, and further detail average excess returns, Sharpe ratios, AUM (US\$ Bi), time series lengths, and the number of funds/ETFs. Panel C, in turn, submits average risk factor returns and Sharpe ratios.

Table 2: Final dataset statistics

Panel A: Mutual Funds							
Description	Mean	Std.	Min	p05	Median	p95	Max
Av. Excess Return (% p.m.)	0.51%	1.30%	-3.90%	-0.39%	0.56%	1.13%	139.84%
S.R. (<i>ex-post</i>)	0.13	0.12	-1.34	-0.08	0.13	0.30	0.69
Av. AUM (US Bi)	0.37	1.54	0.00	0.01	0.06	1.35	47.37
Time Series Length	126	80	30	34	105	292	180
# Total Number of Funds	13,922						
Panel B: Selected ETFs							
Description	Mean	Std.	Min	p05	Median	p95	Max
Av. Excess Return (% p.m.)	0.56%	0.19%	0.13%	0.26%	0.55%	0.94%	0.98%
S.R. (<i>ex-post</i>)	0.12	0.04	0.03	0.04	0.13	0.18	0.19
Av. AUM (US Bi)	6.80	15.77	0.10	0.23	2.51	20.72	137.89
# Total Number of Funds	100						
Panel C: Risk Factors							
Description	Av. Excess Return (% p.m.)		S.R. (<i>ex-post</i>)				
Mkt	0.72%		0.17				
SMB	0.02%		0.01				
HML	-0.12%		-0.05				
WML	0.10%		0.02				

Note: The table displays Statistics for the final dataset. Panel A refers to the mutual fund sample, Panel B to the selected ETF, and Panel C to risk factors. The table includes average excess returns, Sharpe ratios, time series length, as well as the number of funds and ETFs.

As we can see in the figure, both mutual funds and ETFs, as well as excess returns and Sharpe ratio averages are positive, although lower than numbers reported for the market risk factor (Panel C). As Cochrane (2009) points out, the fact that the market risk factor yields higher Sharpe ratios than average fund Sharpe ratios may suggest that said risk factors may be able to spam most funds' excess returns, whereas only a small number of funds would be able to generate alpha. In fact, this result is confirmed by Barras *et al.* (2010) for the period ranging from 1975 to 2006. In Panel B of Table 2, we notice that average ETF Sharpe numbers are closer to average fund Sharpe ratios, thereby implying that funds, on most cases, are not alpha when benchmarked by ETFs. On average, each fund is examined monthly over a 10-year-term, resulting in an appropriate time-series length.

4 Results

4.1 Benchmark potential of ETFs

In this section we show the potential ETFs have to explain fund return distributions. We start by applying the risk factor model to evaluate results, used herein

as our performance guideline. To do so, we look at the proportion of skills across three subsamples: the first one encompasses the entire period ranging from 1980 to 2019; the second brings observations from 1980 to 2004, a timeframe similar to the one examined in the Barras *et al.* (2010) paper; finally, the third and main period used herein covers the years from 2005 to 2019.

Proportions of unskilled, zero-alpha and skilled funds are shown in Table 3, in accordance with the sub-sample that was analyzed.

Table 3: Proportions of zero-alpha, unskilled, and skilled funds

Value	π_-	π_0	π_+
Panel A: Period 01/1980 to 12/2019			
Proportion	37.86%	60.95%	1.19%
Number	5,271	8,485	163
Std.	0.0052	0.0109	0.0083
Panel B: Period 01/1980 to 12/2004			
Proportion	26.29%	73.17%	0.54%
Number	1,494	4,158	30
Std.	0.0077	0.0138	0.0099
Panel C: Period 01/2005 to 12/2019			
Proportion	41.27%	57.56%	1.17%
Number	4,829	6,732	137
Std.	0.0056	0.0116	0.0088

Note: The table displays proportions of zero-alpha, skilled, and unskilled (π_0, π_+, π_-) funds based on the risk factor model for three different time samples. Panel A shows estimated proportions for the entire fund population (13,922 funds) from 1980 to 2019. Panel B submits estimated proportions from the entire sample from 1980 to 2004 (5,682 funds). Panel C exhibits estimated proportions in the sample from 2005 to 2019 (11,700 funds). γ^ and λ^* values are obtained using the selection algorithms devised by (Barras *et al.*, 2010).*

The table replicates specific patterns documented by academic literature, these being: predominance of unskilled and zero-alpha funds, whose joint proportions exceed 98% in all subsamples; and an increase in the volume of unskilled funds over time with the proportion rising from 26.29% in the 1980/2004 period to 41.27% in the 2005/2019 period. Barras *et al.* (2010) also report both results.

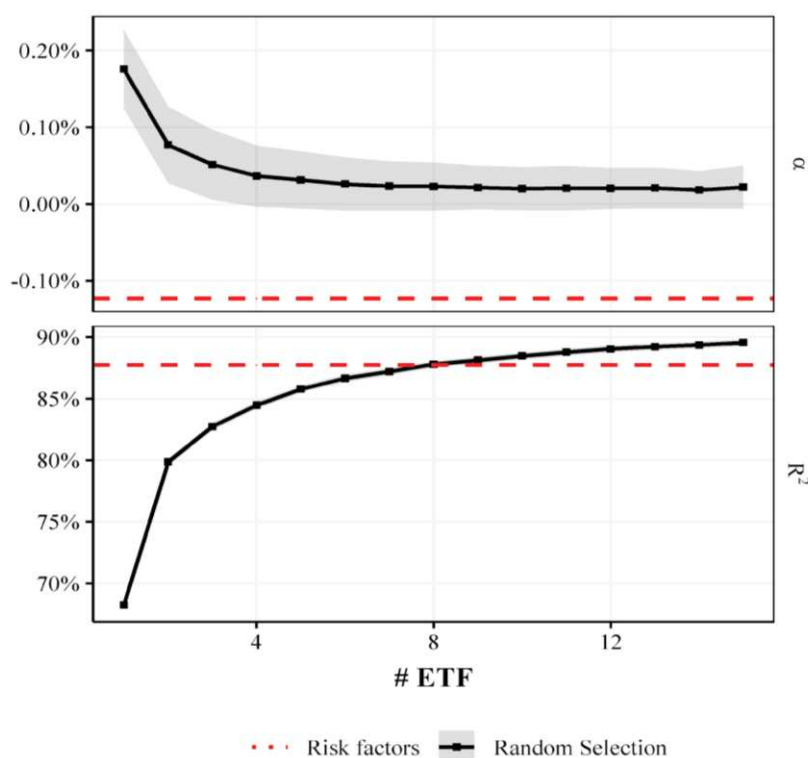
Nevertheless, these conclusions are reached by comparing fund returns with traditional risk factors, these being theoretical portfolios that pose several caveats in regards to their practical elaboration (Frazzini *et al.*, 2012; Novy-Marx & Velikov, 2016).

Under this context, we consider the sensitivity of these results whenever funds are exposed to ETF portfolios instead of traditional risk factors. Using the method detailed in section 2.2, we proceed to randomly pick ETF combinations and assess combinations with 1 to 10 ETFs. First, we define the number of ETFs that we will use as benchmarks,

after which point we choose 500 random combinations, whose average alpha and average adjusted-R² distributions are evaluated.

The results are summarized in Figure 6. The upper panel represents the average alpha according to the number of ETFs used across the 500 random combinations tested, as well as their 95% confidence interval. The lower panel, in turn, shows the adjusted-R², according to the number of ETFs considered. Each point in the figure illustrates an average number of 500 random combinations, whose individual values are calculated from a distribution of 11,700 funds, as detailed in section 2.2. The red dotted lines refer to average alpha risk factors of -0.13% and the average adjusted-R² of 87.7%.

Figure 6: Random ETF Selection Potential vs Risk Factors



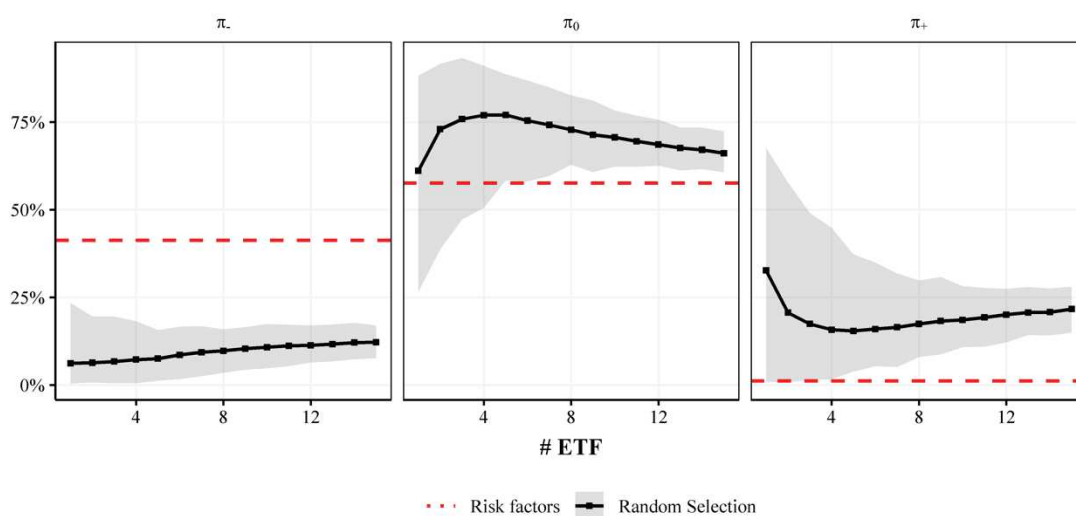
Note: The figure shows the average alpha (α) and the adjusted-R² estimated using the risk factor model based on Random ETF selection, conditional to the number of ETFs selected. The red dotted line represents risk factor estimates, whereas the solid black line denotes estimates based on Random ETF selection: avg. alpha of -0.13% and avg. adjusted-R² of 87.7%.

Figure 6 enables us to note that the average alpha stabilizes when we take into consideration combinations of around 4 ETFs. Furthermore, the value stabilizes at levels of 0.02%, showing that it is approximately 0.15% higher than the average alpha calculated with risk factors. This substantial difference is close to the estimated transaction cost of 0.18% used to devise risk factors, according to Frazzini *et al.* (2012).

Regards to the adjusted-R², the results are also noteworthy since whenever we evaluate combinations of 4 or more ETFs, this measure is also close to the one we obtain with risk factors. When we examine combinations with 8 ETFs, the adjusted-R² reaches average numbers akin to those obtained with risk factors (87.7%), whereas from 4 ETFs onwards the average already exceeds 85%.

Thereafter, we proceed to analyze Figure 7, where we show average proportions of unskilled, zero-alpha and skilled funds, according to the number of ETFs evaluated in random combinations. The solid black line refers to average proportions among random combinations concerning different numbers of ETFs that we analyzed. The gray area delimits intervals from the 2.5th to the 97.5th percentiles of random combinations used. The red dotted line denotes proportions estimated with risk factors, as seen in Panel C of Table 3: 41.27%, 57.56%, and 1.17% for unskilled, zero-alpha and skilled funds, respectively.

Figure 7: Expected skill proportion based on Random ETF Selection



Note: The figure shows average proportions of unskilled, zero-alpha, and skilled funds, according to the number of ETFs tested in the random selection. The solid black line refers to average values of each proportion among random combinations tested pursuant to the number of ETFs analyzed. The gray area delimits intervals from the 2.5th to the 97.5th percentiles. The red dotted line indicates skill proportions estimated using risk factors.

Starting with the proportion of unskilled funds, the figure backs our results previously shown in Figure 6. The proportion of unskilled funds does not correspond to the number obtained with risk factors when we use ETFs. Benchmarking fund returns with ETF combinations entail unskilled fund proportions with stable averages of around 10%, regardless of the number of ETFs used. When we consider risk factors, however, this proportion reaches 41.27%. This behavior can also be explained by the absence of

transaction costs in risk factor returns. Without transaction costs, risk factor averages are potentially higher, thus, the use of such variables as benchmarks signal potentially lower alphas, in addition to a higher unskilled proportion. Such a pattern is consistent with the conclusion found in Figure 6, in which we point out that the average alpha is higher when we use ETFs as a reference as opposed to when we use risk factors.

In contrast, although proportions of unskilled funds are stable, proportions of zero-alpha and skilled funds seem to be sensitive to the number of ETFs analyzed, behaving in a total opposite way. As we look at more and more ETF combinations, proportions of calculated zero-alphas go up, while proportions of estimated skilled funds go down. In truth, we expect the opposite to happen, since all three proportions add up to 100%, and the unskilled fund proportion remains stable. Nevertheless, when we examine intervals from the 2.5th to the 97.5th percentiles, we note that ETF combinations yielding results akin to those obtained with risk factors exist since the gray area intersects with the red dotted line which represents the risk factors. Additionally, in regards to average proportions, it becomes clear that the best results, in order to explain fund performance, are obtained using combinations that range from 3 to 5 ETFs due to the fact that zero-alpha proportions reach their peak in this interval, whereas proportions of skilled funds sink to their lowest point.

Another pattern, shown in Figure 7, refers to the positive trend concerning proportions of skilled funds, which may help to tell us that, as we proceed to test combinations with increasing numbers of ETFs, the proportion of skilled funds will also rise. At first, the pattern may seem unusual, however, one must keep in mind that ETFs are selected randomly, consequently, insofar as we start to consider combinations with greater numbers of ETFs, the number of possibilities expands rapidly, thereby making it difficult to evaluate the same combinations again only with the increase of specific ETFs²⁵.

Our assessment in this section enables us to conclude that ETFs demonstrate a capacity to generate results that come close to those obtained when we compare fund returns to risk factors, although they show inconsistencies concerning the proportion of estimated unskilled funds, which, again, are potentially explained by differences in

²⁵ For instance, when combinations of 4 out of 100 ETF options are drawn, the total number of options is 3,921,225 ($= 100! / ([4! (100-4)!])$). However, when we look at combinations of 6 ETFs, the number of options is 186,087,894,300 ($= 100! / ([8! (100-8)!])$).

transaction costs. On the other hand, when we choose to single out proportions of skilled funds for evaluation, we observe that combinations around 4 ETFs are best suited to generate close results to the ones attained using risk factors.

In summary, we conclude that: *(i)* ETFs can be used as benchmarks to assess fund performance; *(ii)* we should only compare skilled fund proportions since transaction costs suggest differences between proportions of unskilled funds obtained with risk factors and ETFs; *(iii)* finally, the optimal ETF combination should not stray far from the range of 3 to 5 ETFs.

4.2 ETF Selection

Given the results of section 4.1, the following are the dedicated methodologies we employ to identify optimal ETF combinations. Unlike the previous section, this one does not intend to evaluate random ETF combinations, but rather combinations chosen through specific procedures.

These procedures are divided into two approaches. The first one is detailed in section 4.2.1 and selects a set of ETFs aimed at replicating the best possible results concerning proportions of skilled, adjusted-R² and alpha funds. The second approach, in turn, is described in subsection 4.2.2 and uses a procedure to identify a set of ETFs that replicates risk factor dimensions. Both approaches have the same purpose: to select the best benchmarking set for the fund sample. However, the first one addresses this through fund return dimensions, whereas the second takes on an approach based on the perspective of risk factors.

4.2.1 ETF selection based on fund performance

In this section we disclose the results obtained selecting ETFs after we examined fund return dimensions. The approach aims to identify ETF combinations capable of explaining fund return variability (high adjusted-R²); to generate a low proportion of skilled funds (low π_+), and to recommend low average values for risk-adjusted excess returns (low α).

As described in section 2.3, the procedure used consists of selecting ETFs sequentially so that an ETF is picked in each iteration, thereby generating the lowest possible proportion of skilled funds with an average adjusted-R² greater than 85%, as well as higher than the adjusted-R² obtained with the last selected ETF. Therefore, the first

ETF is selected in order to minimize both the proportion and the skilled funds, though conditioned to an average adjusted- R^2 equal to or greater than 85%. We then continue to add further ETFs, always minimizing the proportion of skilled funds, and only considering options that generate an average adjusted- R^2 greater than or equal to the one obtained in the previous iteration. The procedure concludes when we establish that additional ETFs are not capable of reducing the estimated proportion of skilled funds.

A minimum adjusted- R^2 average is required to help avoid regressions that show weak adherence to data, as well as to reduce potential calculations of p-values with volatile distribution for both skilled and unskilled funds. As Andrikogiannopoulou & Papakonstantinou (2019) point out, excessively volatile calculated p-values can lead to significantly high values, even for non-zero-alpha funds. In this case, we see an overestimation of the proportion of zero-alphas, which, in turn, leads to an underestimation of proportions of non-zero alpha funds.

After applying the aforementioned procedure, we proceed to select five ETFs: VV.U; IJT; SLYV; XLY; and XLV. Results concerning performance metrics distribution used in this proposed selection (combining the four benchmarking ETFs), in addition to the estimated proportions of unskilled, zero-alpha and skilled funds, are found in Table 4. With the aim of helping to compare outcomes obtained, and taking into account risk factors and randomly selected ETFs, we also include them in the table. Panel A shows performance distribution metrics, whereas Panel B displays skill proportions.

The results exceed those attained simply by randomly selecting ETFs, considering every metric analyzed. Moreover, they are close to values obtained with risk factors, except for the average alpha, which in spite of being negative, is still approximately 0.11% higher than the estimated risk factors.

Finally, as one can see, the estimated proportion of skilled funds accounts for only 5.73%. Therefore, based on the selected ETF combination, we estimate that 94.27% of both population and funds fail to demonstrate adequate management skills required to generate positive risk-adjusted returns. The result underscores evidence that has already been shown in a wide host of papers, nevertheless, when considering all transaction costs, it is still attained herein. Similar results can be achieved by analyzing funds' gross returns, i.e., before expenses. However, as studies have also shown, there are far too many distortions in the available expense ratio data (Berk & van Binsbergen, 2015, 2017).

Table 4: ETF selection results based on fund performance distribution

Panel A: Performance measures							
Benchmark	Mean	Std.	Min	p05	Median	p95	Max
α							
<i>Risk factors</i>	-0.12%	0.92%	-0.93%	-0.46%	-0.13%	0.13%	0.42%
<i>Random selection</i>	0.03%	1.84%	-0.82%	-0.32%	0.03%	0.26%	0.48%
<i>Proposed Selection (PS)</i>	-0.01%	0.67%	-0.87%	-0.35%	0.00%	0.24%	0.56%
$adjusted - R^2$							
<i>Risk factors</i>	87.72%	16.58%	3.20%	51.94%	92.91%	97.78%	99.08%
<i>Random selection</i>	86.43%	15.65%	4.14%	55.13%	91.29%	95.75%	97.24%
<i>Proposed Selection (PS)</i>	86.15%	16.97%	-1.13%	51.16%	91.41%	97.16%	98.63%
Panel B: Skill distribution							
	π_-		π_0		π_+		
<i>Risk factors</i>	41.28%		57.54%		1.18%		
<i>Random selection</i>	9.67%		71.31%		19.02%		
<i>Proposed Selection (PS)</i>	5.63%		88.64%		5.73%		

Note: Panel A of the table displays the distribution of alpha and adjusted-R² in the funds that we analyzed, in accordance with the evaluated benchmark: Risk factors, Random Selection or Proposed Selection (PS). In regards to random selection, average estimated parameters for each fund are shown, taking into account models with 1 to 15 ETFs. Panel B reports proportions of unskilled, zero-alpha, and skilled funds.

Still, it is worth pointing out that despite some impressive results, they might be impacted by overfitting fund data samples that we analyzed. This stems from the fact that several variables are evaluated to choose the best set of ETFs. Consequently, we also examine out-of-sample results in the following section.

4.2.1.1 Cross-validation

This section employs the same ETF selection procedure as the one described in section 4.1, however, here we proceed to compare both the in-sample and out-of-sample performance. To achieve this, fund populations were randomly divided into two groups: the training-set and the testing-set. The training-set sample, which covers 70% of the fund sample, is used to select ETFs. The testing-set sample, in turn, contains the remaining 30% funds, and its sole purpose is to ensure an overfitting-free performance. To better assess the method performance, we execute repeated cross-validation, wherein we carry out the described procedure 100 times, in addition to evaluating distributions of average alpha, average adjusted-R² and the estimated π_+ .

The results are in Table 5, where we disclose parameter distributions for all 100 cross-validation samples. Furthermore, as a reference parameter, we report the results we obtained with the proposed algorithm selection (PS) for every sample shown in the previous section. It is essential to point out that full sample results are estimated only once, thus, there is only one value for the average α , the average adjusted-R², and π_+ .

Table 5: Cross-validation for ETF selection based on fund performance

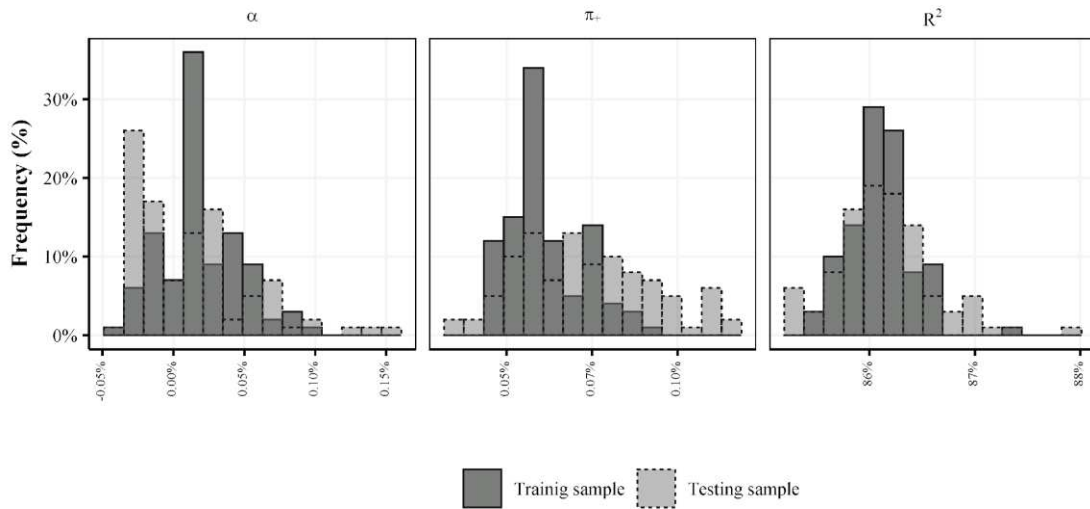
Sample	Full sample	Training	Testing
Average α			
Mean	-0.01%	0.02%	0.02%
Std.	-	0.03%	0.04%
Min	-	-0.04%	-0.04%
p05	-	-0.02%	-0.03%
Median	-	0.02%	0.01%
p95	-	0.07%	0.11%
Max	-	0.09%	0.15%
Average adjusted-R²			
Mean	87.72%	86.07%	86.03%
Std.	-	0.25%	0.41%
Min	-	85.45%	85.08%
p05	-	85.59%	85.30%
Median	-	86.08%	86.04%
p95	-	86.47%	86.76%
Max	-	86.71%	87.09%
π_+			
Mean	5.73%	6.07%	7.38%
Std.	-	1.00%	2.02%
Min	-	4.34%	3.44%
p05	-	4.65%	4.65%
Median	-	5.95%	7.47%
p95	-	7.58%	10.92%
Max	-	8.52%	12.39%

Note: Panel A of the table displays the distribution of the average alpha and the average adjusted-R² throughout the cross-validation samples for the proposed algorithm selection (PS), in accordance with the evaluated sample. We only disclose the mean for the Full sample since the only number available is the estimated result from Table 4 section 4.2.1. For Training-sets and Testing-sets, numbers shown relate to average α and average adjusted-R² parameters estimated throughout all 100 cross-validation samples.

As we can see, results obtained with training-sets and testing-sets are close, as well as results reported for the entire fund sample disclosed in the previous section. The alphas' training and testing distributions are very similar to each other, and both present estimated averages around 0.02% close to the full sample result of -0.01%. Besides that, the training and testing samples have adjusted-R² averages around 86%, which is close to the 88% full sample estimated value. The standard deviations for the alphas and adjusted-R² are low for the training and testing samples. Proportions of skilled funds have higher variabilities, with average values slightly above those calculated in the full sample. The training sample has an estimated proportion of 6.07%, while the testing sample records 7.38%. Nevertheless, both numbers are close to the full sample result of 5.73%.

Figure 8 allows us to compare the distribution of estimated metrics for both training-sets and testing-sets across all simulations performed. Again, it is possible to verify that distribution percentages are close, signaling that the procedure has a low overfitting risk, which helps to underscore the results we have already disclosed.

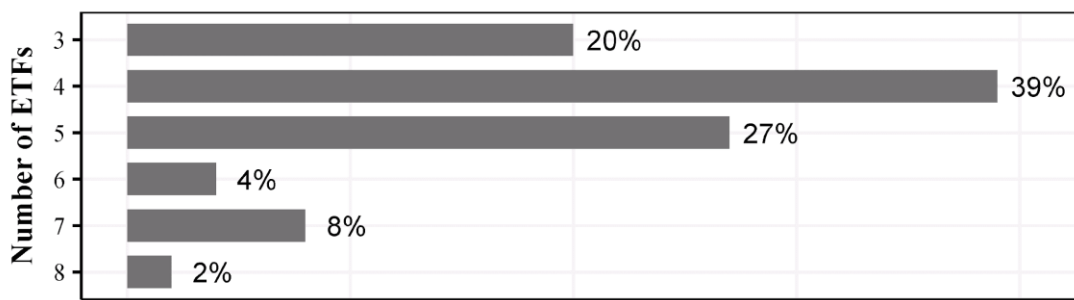
Figure 8: Measure distributions among training-sets and testing-sets



Note: The figure displays alpha distributions (α), the proportion of skilled funds (π_+), and the adjusted- R^2 obtained with the cross-validation procedure. Histograms in dark gray and solid lines denote the distribution of estimated parameters in the training sample, while histograms in light gray and the dotted line represent distributions in the testing sample.

It is also worth evaluating the sensitivity of the results in regards to the number of ETFs selected. When we look at the entire fund sample, we select five ETFs reinforcing the quantity indicated in the random ETF selection analysis described in section 4.1. Likewise, Figure 9, which illustrates how the selected ETFs are distributed, shows that nearly 86% of cases are performed with a number of ETFs ranging from 3 to 5. Again, close to the pattern observed with the entire data sample.

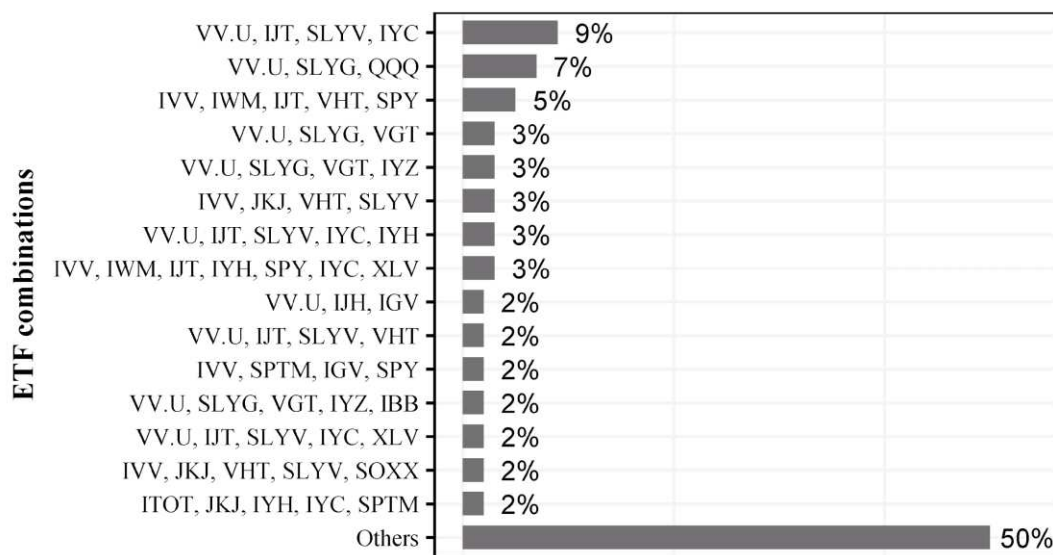
Figure 9: Cross-validation distribution for the number of selected ETFs



Note: The figure displays the distribution of the number of ETFs selected with the cross-validation procedure. The vertical axis refers to the number of ETFs selected, while the horizontal axis represents their frequency.

Finally, Figure 10 shows how the selected ETF combinations are distributed. Combinations are arranged from lowest to highest in terms of frequency. Furthermore, combinations picked only once are grouped in the “Other” category.

Figure 10: ETF combination frequency in cross-validation



Note: The table shows the frequency by which each ETF combination was selected using the cross-validation procedure. The vertical axis depicts ETF combinations, whereas the horizontal axis illustrates the frequency in which they occur. Combinations that occur only once are grouped in the "Other" category.

As we can see, the set of selected ETFs does not exhibit a pattern. In other words, the procedure does not seem to suggest that any optimal ETF combination exists, although there are several possibilities. This becomes clear when we take into account the fact that most combinations are selected once, at most, and that the most common combination takes place in approximately only 9% of the cases.

4.2.2 ETF selection based on risk factors

In this section, we employ specific procedures intending to select a set of ETFs capable of mimicking risk factor returns. In doing so, our goal is to attempt to find a set of ETFs that can replicate results when fund returns are exposed to risk factors.

The first tested procedure is based on the Lasso methodology and considers the set of ETFs as potential explanatory variables for risk factors. Thus, we apply a Lasso methodology for each risk factor to pinpoint sets of ETFs capable of explaining risk factor variability. The methodology’s hyperparameters are set following the five-fold cross-

validation method. ETFs selected for each risk factor are found in Table 6. The lower row of the table reports the adjusted-R² obtained after regressing risk factors by the selected ETFs.

Table 6: Selected ETFs by Lasso

ETF	Description	Selected			
		Mkt	SMB	HML	WML
XLF	Blend Funds	x		x	x
IYM	Blend Materials	x			
IYG	Blend Funds			x	
IWO	Small-Cap Growth	x	x		
XNTK	Growth Technology	x			
SOXX	Blend Technology	x			
adjusted -R²		95.5%	74.5%	53.4%	33.9%

Note: The table lists ETFs selected using the Lasso method. Each row refers to a selected ETF. The first column represents the name of the ETF. The second row brings a brief description of the asset investment guideline. The other columns show which risk factor is linked to each ETF. Finally, the last row reports the adjusted-R² attained from the regression between the selected ETFs and each risk factor.

As we can see, the procedure selects five ETFs to explain the Mkt risk factor variability, however, two of them were also selected through HML, WML and SMB. The HML factor requires two ETFs to be chosen. Finally, SMB and WML call for the selection of one additional ETF, each. In all, there are six ETFs selected, whereas in regards to the adjusted-R², they range from 95.5% to 33.9%.

The second selection procedure is based on the forward selection methodology. More specifically, said procedure consists of selecting a set of ETFs for each risk factor to obtain the highest possible adjusted-R². This approach picks one ETF for each risk factor in each iteration. Unless the same ETF is selected to explain different risk factors, the procedure selects a set with four ETFs in each new iteration.

Only results obtained with, at most, 2 ETFs for each risk factor were evaluated. Selected ETFs are shown in Table 7, where Forward 1 and Forward 2 submit, respectively, results obtained when one to two ETFs are selected for each risk factor.

Results show that ETFs selected for each risk factor do not intersect, thus, we picked an overall number of four ETFs in the first iteration (Forward 1), plus an additional four in the second iteration (Forward 2). The adjusted-R² ranges from 99.6% to 33.2% when we use one ETF, and from 99.6% to 45.5% when we select two ETFs.

Final results obtained using the procedure described are found in Table 8, in addition to results attained with risk factors and random ETF selection. The table is split into two panels. Panel A shows the distribution of alpha and adjusted-R² for each set of

benchmarks. Panel B, in turn, reports proportions of unskilled, zero-alpha, and skilled funds.

Table 7: Selected ETFs by Forward Selection

ETF	Benchmark	Selected			
		Mkt	SMB	HML	WML
Panel A: Forward 1					
XLF	Blend Funds				x
IWM	Small-Cap Blend		x		
SPTM	Broad Market Blend	x			
PEY	Large-Cap Blend			x	
adjusted -R²		99.6%	94,5%	57.4%	33.2%
Panel B: Forward 2					
XLF	Blend Funds				x
IWM	Small-Cap Blend		x		
SPTM	Broad Market Blend	x			
PEY	Large-Cap Blend			x	
SPTM	Broad Market Blend	x			
FVL	Broad Market Blend				x
JKE	Large-Cap Growth			x	
PEY	Large-Cap Blend			x	
SPTM	Broad Market Blend	x			
adjusted -R²		99.6%	94.7%	80.6%	45.5%

Note: The table lists ETFs selected with the Forward Selection method. Each row refers to a selected ETF. The first column represents the name of the ETF. The second row brings a brief description of the asset investment guideline. The other columns report which risk factor is linked to each ETF. The last row shows the adjusted-R² from the regression attained between the selected ETFs and each risk factor. Panel A displays the selected ETFs with one method iteration, while Panel B depicts ETFs selected using two iterations.

When we begin looking at Lasso selection results, it is noticeable that this procedure entails the worst results, these being even lower than those we achieved with the random ETF selection. Applying it generates an average alpha of 0.13%, an average adjusted-R² of 84.98%, and a proportion of skilled funds corresponding to 39.37%. On the other hand, both Forward 1 and 2 selection methods deliver better results than those attained using the random ETF selection. In regards to the calculated average alpha for these funds, it is worth pointing out that the average of both approaches comes close to zero (0.02% and 0.01%). We also see satisfactory results for the adjusted-R², which are even closer to the outcome we obtained with the risk factors. Finally, when it comes to the estimated proportion of skilled funds, said ratio varies close to 11%.

The procedures applied suggest that it is increasingly possible to select ETFs that are able to explain returns posted by actively managed investment funds. However, although these results exceed those obtained with random ETF selection, the estimated proportion of skilled funds remains substantially higher than the proportion estimated using risk factors. It is also worth noting that said procedures are not impacted by potential

data snooping since we select ETFs here by analyzing risk factor dimensions, instead of directly from the funds.

Table 8: Selection results based on risk factors

Panel A: Performance measures							
Benchmark	Mean	Std.	Min	p05	Median	p95	Max
α							
<i>Risk factors</i>	-0.12%	0.92%	-0.93%	-0.46%	-0.13%	0.13%	0.42%
<i>Random selection</i>	0.03%	1.84%	-0.82%	-0.32%	0.03%	0.26%	0.48%
<i>Lasso</i>	0.13%	0.90%	-0.81%	-0.29%	0.14%	0.44%	0.71%
<i>Forward 1</i>	0.02%	1.35%	-0.83%	-0.35%	0.02%	0.29%	0.58%
<i>Forward 2</i>	0.01%	1.93%	-0.88%	-0.36%	0.01%	0.27%	0.57%
R^2							
<i>Risk factors</i>	87.72%	16.58%	3.20%	51.94%	92.91%	97.78%	99.08%
<i>Random selection</i>	86.43%	15.65%	4.14%	55.13%	91.29%	95.75%	97.24%
<i>Lasso</i>	84.98%	16.02%	0.31%	52.88%	89.94%	95.63%	97.70%
<i>Forward 1</i>	87.38%	16.61%	1.37%	53.46%	92.56%	97.53%	98.80%
<i>Forward 2</i>	88.79%	16.16%	2.37%	56.35%	93.88%	98.00%	99.08%
Panel B: Skill distribution							
	π_-		π_0		π_+		
<i>Risk factors</i>	41.28%		57.54%		1.18%		
<i>Random selection</i>	9.67%		71.31%		19.02%		
<i>Lasso</i>	4.59%		56.04%		39.37%		
<i>Forward 1</i>	4.36%		84.39%		11.25%		
<i>Forward 2</i>	8.48%		79.89%		11.63%		

Note: Panel A of the table displays the distribution of alpha and adjusted-R² in the funds we analyzed, in accordance with the evaluated benchmark. In regards to random selection, we report the average parameter estimated for each fund, considering models with 1 to 15 ETFs. Panel B, in turn, shows proportions of unskilled, zero-alpha, and skilled funds.

4.2.3 Growth impact on the number of ETFs

As illustrated in section 3, recent years have seen a substantial increase in the volume of ETFs available for investors. Previous sections of this paper examined data from 2005 with 100 ETFs. Nevertheless, when we analyze the database starting from 2009, we can evaluate 299 ETFs in all. That being said, this section describes our results after analyzing this period that features a greater number of ETFs.

More specifically, we examine all prior results in this section, i.e., risk factors, random selection, selection based on fund returns, and selection based on risk factor returns. Nevertheless, we do so by looking at the 2009 to 2019 period, which includes a pool of 299 ETFs.

The results obtained are shown in Table 9. Panel A displays parameters of the calculated alpha distribution for each benchmark set, whereas Panel B illustrates the adjusted-R² distribution. Finally, Panel C brings each fund's skill ratios, also in accordance with the analyzed benchmark set.

Table 9: Results obtained with larger ETF samples

Panel A: Performance measures							
Benchmark	Mean	Std.	Min	p05	Median	p95	Max
α							
<i>Risk factors</i>	-0.15%	0.21%	-0.97%	-0.48%	-0.13%	0.12%	0.38%
<i>Random selection</i>	0.03%	0.20%	-0.73%	-0.29%	0.04%	0.29%	0.52%
<i>Lasso</i>	-0.07%	0.23%	-0.95%	-0.42%	-0.07%	0.25%	0.61%
<i>Forward 1</i>	0.01%	0.21%	-0.89%	-0.32%	0.02%	0.29%	0.55%
<i>Forward 2</i>	0.02%	0.22%	-0.79%	-0.29%	0.04%	0.32%	0.61%
<i>PS</i>	-0.01%	0.20%	-0.79%	-0.32%	0.00%	0.27%	0.58%
R^2							
<i>Risk factors</i>	87.48%	17.00%	3.23%	49.16%	93.03%	97.89%	99.12%
<i>Random selection</i>	82.61%	16.11%	2.39%	47.99%	87.69%	93.10%	94.95%
<i>Lasso</i>	86.98%	16.57%	1.77%	52.73%	92.27%	97.68%	98.74%
<i>Forward 1</i>	86.44%	16.95%	1.86%	50.16%	91.80%	97.33%	98.67%
<i>Forward 2</i>	87.18%	16.57%	2.11%	51.47%	92.52%	97.16%	98.24%
<i>PS</i>	85.57%	17.40%	-0.36%	47.52%	91.14%	97.00%	98.47%
Panel B: Skill distribution							
	π_-		π_0		π_+		
<i>Risk factors</i>	44.94%		52.25%		2.81%		
<i>Random selection</i>	13.19%		65.43%		21.38%		
<i>Lasso</i>	19.51%		78.48%		2.01%		
<i>Forward 1</i>	4.27%		86.04%		9.69%		
<i>Forward 2</i>	5.79%		79.51%		14.70%		
<i>PS</i>	1.19%		98.79%		0.02%		

Note: The table refers to results obtained from observations made in 2009 and 2019, including a sample of 299 ETFs. Panel A of the table reports the distribution of alpha and adjusted- R^2 in the funds we analyzed, in accordance with the evaluated benchmark. In regards to random selection, the table shows the average parameter estimated for each fund, considering models with 1 to 15 ETFs. Panel B displays proportions of unskilled, zero-alpha, and skilled funds.

As we can see, results attained with the new ETF set time-sample are not different from those observed earlier. According to risk factors assessed, we estimate an average alpha of -0.15%. Furthermore, we conclude that 97% of the mutual fund industry fails to generate value. When randomly selecting ETFs, we reach an average alpha of 0.03% and an unskilled fund proportion of 79%. To conclude, we can say that the PS method delivers reliable results again with an average alpha of -0.01%, while signaling that 99.98% of funds fail to generate value for their investors.

5 Conclusion

The number of ETFs has grown massively over the past years. In this paper, we examine this fact to better understand skill distribution relating to active fund performance. We have shown how several combinations within the ETF population are suited to be used as benchmarks to evaluate active management funds. Results obtained in regards to the estimated proportion of skilled managers are similar to those estimated with traditional risk factors. Nevertheless, ETFs are a better fit for the task since, unlike risk factors, they are legitimate financial instruments that are actually traded in the market.

Therefore, their observed performance takes into account transactional costs involved in their operation.

After analyzing random selections of ETF combinations, we are able to verify that: *(i)* ETFs can be used as benchmarks to assess fund performance; *(ii)* we should only compare skilled fund proportions since transaction costs suggest differences between proportions of unskilled funds obtained with risk factors and ETFs; *(iii)* finally, we find that the optimal ETF combination should not stray far from the range of 3 to 5 ETFs.

Nonetheless, we also find that several ETF combinations lead to substantial differences in regards to observed behaviors in the active fund population. Thus, this proves the need for a guided method to select ETFs featuring the appropriate traits required to evaluate funds.

Under this context, we proceeded to test a set of selection algorithms, establishing that it is possible to find suitable combinations. We have also examined cross-validation-based algorithm performances, which led us to the finding that, based on an assessed set of approximately 5 ETFs, 95% of the active mutual fund industry does not have neither the necessary skills nor the ability to generate value for their investors.

Appendix

A1. Procedure to compute alpha p-value for each fund (\hat{p}_i)

For each estimation attained by model (1) ($i = (1, \dots, I)$), we compute the t-statistics for test $H_i^0: \alpha_i = 0$ against $H_i^1: \alpha_i \neq 0$ by $\hat{t}_i = \frac{\hat{\alpha}_i}{\hat{\sigma}_{\alpha_i}}$, where $\hat{\sigma}_{\alpha_i}$ is a HAC estimator (Newey & West, 1987). As Kosowski *et al.* (2006) point out, t-statistic distribution in mutual funds is usually not normal, thus, we employ a bootstrap procedure to compute the p-value for the test. Following the guidelines set by Barras *et al.* (2010), we generate B bootstrap samples²⁶ for each fund return ($\{r_{i,t}^b\}_{b=1}^B$) and compute a series of t-statistics ($\{\hat{t}_i^b\}_{b=1}^B$). The estimated p-value for each fund can be expressed as:

$$\hat{p}_i = 2\min\left(\frac{1}{B}\sum_{q=1}^B \mathbb{I}(\hat{t}_i^q > \hat{t}_i), \frac{1}{B}\sum_{q=1}^B \mathbb{I}(\hat{t}_i^q < \hat{t}_i)\right) \quad (23)$$

where \mathbb{I} is an indicator function that is equal to 1 if $\hat{t}_i^b > \hat{t}_i$ or, otherwise, to zero.

A2. Procedure to compute optimal lambda from the data (λ^*)

Using Barras *et al.* (2010) as our standard again, we set a grid for λ so that $\lambda = (\lambda_1, \dots, \lambda_K)$ ²⁷. For each value of λ_k so that $\lambda_k \in \lambda$, we calculate $\hat{\pi}_0(\lambda_k) = \frac{1}{I} \frac{\sum_{i=1}^I \mathbb{I}(\hat{p}_i > \lambda_k)}{(1-\lambda_k)}$, where \mathbb{I} is an indicator that is equal to 1 if $\hat{p}_i > \lambda_k$ or, otherwise, to zero. Next, we generate B p-value bootstrap samples for each fund i ($\{\hat{p}_i^b\}_{b=1}^B$) and compute $\{\hat{\pi}_0^b(\lambda_k)\}_{b=1}^B$ for each λ_k ²⁸. The mean square error of $\hat{\pi}_0(\lambda_k)$ is as follows:

$$\widehat{MSE}(\lambda_k) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\pi}_0^b(\lambda_k) - \min_{\lambda \in \lambda} \hat{\pi}_0(\lambda) \right)^2 \quad (24)$$

The optimal lambda (λ^*) is defined as:

$$\lambda^* = \operatorname{argmin}_{\lambda \in \lambda} \widehat{MSE}(\lambda) \quad (25)$$

²⁶ We set $B = 1,000$, in accordance with the Barras *et al.* (2010) procedure.

²⁷ We follow the Barras *et al.* (2010) specifications and set $\lambda = (0.30, 0.32, \dots, 0.70)$ and $B = 1,000$.

²⁸ $\hat{\pi}_0^b(\lambda_k) = \frac{1}{I} \frac{\sum_{i=1}^I \mathbb{I}(\hat{p}_i^b > \lambda_k)}{(1-\lambda_k)}$, where \mathbb{I} is a indicate that is equal to 1 if $\hat{p}_i^b > \lambda_k$ or, otherwise, to zero.

A3. Procedure to compute optimal gamma from the data ($\boldsymbol{\gamma}^*$)

Akin to the procedure we used to set λ^* , we start by defining a grid $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)^{29}$. First, we compute $\hat{\pi}_+(\gamma_k) = \hat{S}_{\gamma_k}^+ - \hat{\pi}_0(\lambda^*) \frac{\gamma_k}{2}$ for each value of γ_k ($\lambda_k \in \lambda$). Second, we generate B p-value bootstrap samples for each fund i ($\{\hat{p}_i^b\}_{b=1}^B$) and calculate $\{\hat{\pi}_+(\boldsymbol{\gamma})\}_{b=1}^B$ ³⁰ for each γ_k . The mean square error of $\hat{\pi}_+(\gamma_k)$ is:

$$\widehat{MSE}(\gamma_k) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\pi}_+^b(\gamma_k) - \max_{\boldsymbol{\gamma} \in \boldsymbol{\gamma}} \hat{\pi}_+(\boldsymbol{\gamma}) \right)^2 \quad (26)$$

The optimal gamma (γ_*) is defined as:

$$\gamma_* = \operatorname{argmin}_{\boldsymbol{\gamma} \in \boldsymbol{\gamma}} \widehat{MSE}(\boldsymbol{\gamma}) \quad (27)$$

A4. Standard Deviation Estimators

In accordance with Barras *et al.* (2010), we assume that $\hat{\pi}$ estimators are stochastic processes indexed by $\boldsymbol{\gamma}$ and λ that converge to a Normal distribution if $I \rightarrow \infty$. Hence, $\hat{\sigma}_{\hat{\pi}_0}$ can be estimated as³¹:

$$\hat{\sigma}_{\hat{\pi}_0} = \left(\frac{(\sum_{i=1}^I \mathbb{I}(\hat{p}_i > \lambda^*)) * (I - \sum_{i=1}^I \mathbb{I}(\hat{p}_i > \lambda^*))}{I^3(1 - \lambda^*)^2} \right)^{\frac{1}{2}} \quad (28)$$

If we estimate $\hat{\pi}_+$ by (6) and $\hat{\pi}_-$ by (7), $\hat{\sigma}_{\hat{\pi}_+}$ and $\hat{\sigma}_{\hat{\pi}_-}$ are given by:

$$\hat{\sigma}_{\hat{\pi}_+} = \left(\frac{(\hat{S}_{\gamma^*}^+ (1 - \hat{S}_{\gamma^*}^+))}{I} + \frac{(\gamma^* \hat{\sigma}_{\hat{\pi}_0})^2}{4} + \frac{(\gamma^* \hat{S}_{\gamma^*}^+)}{(1 - \lambda^*)} + \frac{(\sum_{i=1}^I \mathbb{I}(\hat{p}_i > \lambda^*))}{I^2} \right)^{\frac{1}{2}} \quad (29)$$

$$\hat{\sigma}_{\hat{\pi}_-} = \left(\hat{\sigma}_{\hat{\pi}_0}^2 + \hat{\sigma}_{\hat{\pi}_+}^2 - \frac{2\hat{S}_{\gamma^*}^+ (\sum_{i=1}^I \mathbb{I}(\hat{p}_i > \lambda^*))}{(1 - \lambda^*) I^2} - \gamma^* \hat{\sigma}_{\hat{\pi}_0}^2 \right)^{\frac{1}{2}} \quad (30)$$

If we use (6) to calculate $\hat{\pi}_-$ and (7) to estimate $\hat{\pi}_+$ in order to obtain $\hat{\sigma}_{\hat{\pi}_+}$ and $\hat{\sigma}_{\hat{\pi}_-}$ expressions, we merely change subscriptions $+$ to $-$, and $-$ to $+$, for equations (29) and (30).

²⁹ Again, we set $\boldsymbol{\gamma} = (0.05, 0.10, \dots, 0.50)$ and $B = 1,000$ conforming Barras *et al.* (2010).

³⁰ $\hat{\pi}_+^b(\boldsymbol{\gamma}) = \frac{\sum_{i=1}^I \mathbb{I}(\hat{p}_i^b > \boldsymbol{\gamma})}{I} - \hat{\pi}_0^b(\lambda^*) \frac{\boldsymbol{\gamma}}{2}$, where \mathbb{I} is an indicator that is equal to 1 if $\hat{p}_i^b > \boldsymbol{\gamma}$ or, otherwise, to zero.

³¹ See Proposition 3.2 from Genovese & Wasserman (2004).

References

- Alexander, C., & Barbosa, A. (2008). Hedging index exchange traded funds. *Journal of Banking & Finance*, 32(2), 326–337.
- Andrikogiannopoulou, A., & Papakonstantinou, F. (2019). Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *The Journal of Finance*, 74(5), 2667–2688.
- Ardia, D., & Boudt, K. (2018). The peer performance of hedge funds. *Cahier de recherche/Working Paper*, 13, 29.
- Barras, L., Scaillet, O., & Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The journal of finance*, 65(1), 179–216.
- Berk, J. B., & van Binsbergen, J. H. (2015). Measuring skill in the mutual fund industry. *Journal of Financial Economics*, 118(1), 1–20. <https://doi.org/10.1016/j.jfineco.2015.05.002>
- Berk, J. B., & van Binsbergen, J. H. (2017). Mutual Funds in Equilibrium. *Annual Review of Financial Economics*, 9(1), 147–167. <https://doi.org/10.1146/annurev-financial-110716-032454>
- Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance*, 57–82. Recuperado de <https://doi.org/10.1111/j.1540-6261.1997.tb03808.x>
- Chen, Y., Cliff, M., & Zhao, H. (2017). Hedge funds: The good, the bad, and the lucky. *Journal of Financial and Quantitative Analysis*, 52(3), 1081–1109.
- Cochrane, J. H. (2009). *Asset pricing: Revised edition*. Princeton university press.
- Crane, A. D., & Crotty, K. (2018). Passive versus active fund performance: do index funds have skill? *Journal of Financial and Quantitative Analysis*, 53(1), 33–64. <https://doi.org/10.1017/S0022109017000904>
- Criton, G., & Scaillet, O. (2014). Hedge fund managers: Luck and dynamic assessment. *Bankers, Markets & Investors*, (129), 28–38.
- CRSP. (2012). Survivos-bias-free US mutual fund guide.
- Fama, E. F. (1998). Determining the number of priced state variables in the ICAPM. *Journal of Financial and Quantitative Analysis*, 33(2), 217–231.
- Fama, E. F., & French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *The journal of finance*, 65(5), 1915–1947.
- Frazzini, A., Israel, R., & Moskowitz, T. J. (2012). Trading costs of asset pricing anomalies. *Fama-Miller working paper*, (14), 5–14.
- Franzoni, F., & Schmalz, M. C. (2017). Fund flows and market states. *Review of Financial Studies*, 30(8), 2621–2673. <https://doi.org/10.1093/rfs/hhx015>
- Genovese, C., & Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3), 1035–1061.
- Harvey, C. R., & Liu, Y. (2019). Cross-sectional alpha dispersion and performance evaluation. *Journal of Financial Economics*, 134(2), 273–296.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when

the alternative hypothesis is true. *Biometrics*, 11–22.

- Kosowski, R., Timmermann, A., Wermers, R., & White, H. (2006). Can mutual fund “stars” really pick stocks? New evidence from a bootstrap analysis. *The Journal of finance*, 61(6), 2551–2595.
- Lettau, M., & Madhavan, A. (2018). Exchange-traded funds 101 for economists. *Journal of Economic Perspectives*, 32(1), 135–154.
- Malkiel, B. G. (1995). Returns from investing in equity mutual funds 1971 to 1991. *The Journal of finance*, 50(2), 549–572.
- Murdoch, D. J., Tsai, Y.-L. L., & Adcock, J. (2008). P-values are random variables. *The American Statistician*, 62(3), 242–245. <https://doi.org/10.1198/000313008X332421>
- Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777–787.
- Novy-Marx, R., & Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies*, 29(1), 104–147. <https://doi.org/10.1093/rfs/hhv063>
- Pástor, L., & Stambaugh, R. F. (2002). Investing in equity mutual funds. *Journal of Financial Economics*, 63(3), 351–380.
- Patton, A. J., & Ramadorai, T. (2013). On the high-frequency dynamics of hedge fund risk exposures. *The Journal of Finance*, 68(2), 597–635.
- Poterba, J. M., & Shoven, J. B. (2002). Exchange-traded funds: A new investment option for taxable investors. *American Economic Review*, 92(2), 422–427. <https://doi.org/10.1257/000282802320191732>
- Sharpe, William F. "Asset allocation: Management style and performance measurement." *Journal of portfolio Management* 18.2 (1992): 7-19
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 479–498.