Evaluating Fund Manager Skill: A Mixture Model Approach

Colin Swaney
University of Iowa

January 10, 2016

Abstract

This paper investigates the distribution of skill in the actively managed equity mutual fund industry using a recently proposed model of fund performance. Our approach to estimation is flexible, intuitive, and gives a complete description of the distribution of skill. We provide new estimates of the size and average performance of the population of skilled active managers. Our results show that the distribution of fund performance is consistent with a population of skilled managers, but that these managers cannot be reliably identified based on past performance alone. Using a simple simulation exercise we show that our recommended estimation method is expected to be accurate under reasonable assumptions about the distribution of fund skill.

1 Introduction

The topic of this paper is evaluating the skill of actively managed equity mutual funds. There are two questions that are of primary interest. First, if we consider the population of mutual funds as a whole, is there any evidence that some funds are more skilled than others? In particular, are there any funds that we expect to produce returns that more than compensate their expenses? Second, supposing that such funds exist, can we identify these funds well enough to capture the value they provide? These questions are of obvious importance to investors choosing between actively and passively managed mutual fund investments.

The literature related to these questions dates back at least as far as [Jensen, 1968]. The early evidence indicated that there is no value to investing in the average actively managed fund. Subsequent studies have focused on the predictability of returns. In an influential paper, [Carhart, 1997] demonstrated that the existence of so-called persistence–the existence of funds that appear to consistently generate value–can be accounted for by measuring exposures to a "momentum" risk-factor. [Baks et al., 2001] points out that from a Bayesian perspective an investor's prior beliefs about the distribution of skill play an important part in determining his or her allocation of wealth to active managers.

The main problem in assessing the large population of mutual funds is differentiating skill from luck. Standard critical values are inappropriate measures of significance when performing multiple hypothesis tests because many of the tests are expected to reject the null hypothesis by chance. In the case of actively managed equity mutual funds the empirical distribution of performance is in fact close to normally distributed, consistent with the hypothesis that all differences in performance are due solely to luck. [Kosowski et al., 2006] addresses this issue using a bootstrapping procedure. Their evidence suggests that the empirical distribution of fund performance is unlikely to be observed through luck alone. Thus they conclude that skilled funds exist, and that their performance exhibits persistence relative to their proposed criteria. More recently, [Barras et al., 2010] attempts to account for lucky outcomes using a method introduced by [Storey, 2002] in the context of biostatistical analyses. Their study finds evidence of persistent skill over short time-horizons, but this evidence is confined to an early sub-sample.

This paper builds on the framework suggested by [Barras et al., 2010]. Our primary insight is that, if taken seriously, their basic framework (which is routinely employed in physical sciences) suggests an alternative estimation technique that is well-known throughout the machine learning community. The primary advantage of the method we propose is that it allows us to more fully describe the distribution of fund performance, which is of importance to investors ([Baks et al., 2001]). Unlike [Barras et al., 2010], this method leads to a direct and intuitive classification scheme, which we use to assess the skill of individual funds. In addition, the method allows for additional flexibility in that it permits the possibility that the majority of funds may generate returns that do not justify their expenses. From a practical perspective this alternative method is preferable for this class of problems because it is much simpler to implement, and standard implementations are freely available in R and Python .

Our first set of experiments focuses on estimation. In this section we employ our methodology to provide a description of the universe of actively managed equity mutual funds in existence during our sample period. We provide estimates of the full set of parameters used in our model of that universe. This main set of estimates shows that the universe of funds under consideration is consistent with a distribution in which a majority (more than 80%) of funds generate slightly negative alpha, and minority populations generating significantly positive and negative alpha, which we label as *Good* and *Bad* performers. These poor and good performing populations are roughly equivalent in size, but the mean performance of the poor performing population is roughy twice that of the good performing population. Unlike [Barras et al., 2010], we find little difference between the distribution of long-run performance and the distribution of performance measured over a shorter time horizon.

From a machine learning perspective our first set of results constitute "unsupervised learning": we have no way of directly testing the quality of description. In the next section we recommend a classification system based on the unsupervised results and quantify its performance in terms of the time series of returns in generates. In order for our strategy to work our classifications must be *stable* over time. In fact, we find limited evidence for persistence of fund performance: funds are unlikely to be classified as *Good* (*Poor*) in years that are close together. In fact, when we analyze the performance of our investment strategies we find that portfolios of good (poor) performers do generate positive (negative) alphas during our sample period, but the performance of these portfolios is quite similar to naive portfolios that invest in the top (bottom) decile of performers over a give period. The primary advantage of our classification system is that it frequently find no *Good* funds.

In our final experiment we address the validity of our methodology in the context of mutual fund analysis by performing a simple simulation exercise. Our results show that the proposed estimation procedure is expected to to produce accurate estimates for plausible distributions. Our simulation also presents an interesting finding: if we accept the model of [Barras et al., 2010], and assume that the estimates provided by those authors are accurate, then our main findings are extremely unlikely.

The remainder of the paper is organized as follows. In Section 2 we explain our framework and describe our proposed estimation procedure. Section 3 describes our data sample. Our main findings are in Section 4, and Section 5 concludes.

2 Evaluating Skill

Our strategy to answer these questions involves two parts. First, we propose a simple statistical model of the distribution of fund skill. Our model assumes the existence of three distinct sub-populations of funds possessing varying degrees of skill: poor, average, and good funds. Second, we use a well-known unsupervised learning method–the Expectation Maximization (EM) algorithm–to estimate the statistical model. This method provides a complete description of the distribution of skill, and leads to a simple and intuitive classification scheme that we use to evaluate individual funds. The details of the statistical model and proposed learning method are described below.

2.1 Framework for evaluating skill

We take for our setting an investor who evaluates individual mutual funds according to each fund's estimated alpha with respect to a given asset pricing model. We assume that the alphas come from one of three sub-populations of funds, which possess increasing levels of skill, and are thus expected to achieve increasing risk-adjusted returns. The investor observes the alpha generated by each fund over some time period, but not the sub-population to which the fund belongs. Her first goal is to discover the proportion of funds that belong to each sub-population. Her second goal is to predict the unobserved class of each fund. It's possible for our investor to perform well on the first task, and yet poorly at the second. This could happen if, for example, the number of funds in a particular sub-population is small, or if skill is conditional on unobserved, time-varying parameters. From an investment stand point the first task may be important even if the later proves difficult.

The specific model we have in mind is a mixture model in which each skill level has a normal distribution (i.e., we consider a mixture of normals distribution). The model is depicted in Figure 1. With (unconditional) probability π_j an observation α_i is drawn from the j-th skill class. The observations of alpha from the j-th class are distributed according to a normal distribution with mean μ_j and standard deviation σ_j. The investor possesses no knowledge of the skill level of each observation (denoted by z_i), and therefore is unable to directly estimate the parameters μ_j, σ_j, and π_j $(j = 1, 2, 3)$. In terms of Figure 1, she only observes the full distribution (the solid line), but she believes that the underlying populations exist and attempts to learn the parameters that describe them.

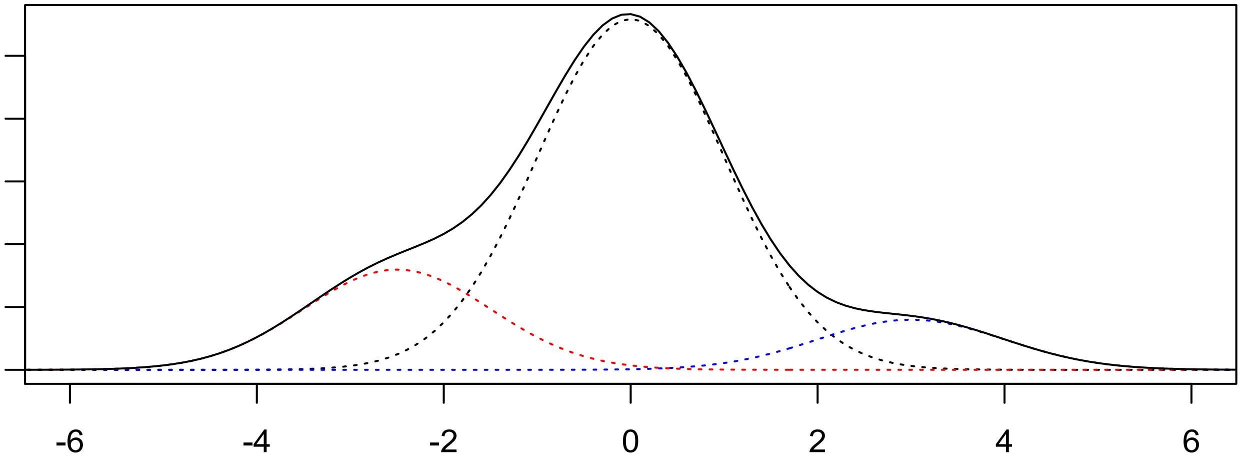


Figure 1: An example mixture-of-normals density. The solid line shows the density of the overall distribution; dashed lines show the densities of the sub-populations.

2.2 Estimation procedure

We use the Expectation-Maximization (EM) algorithm to estimate the parameters of the model. The EM algorithm is a numerical method used to obtain maximum-likelihood estimates of parameters in statistical models that contain latent variables. For example, medical image reconstruction may involve the identification of unobserved tissue types from MRI data. The addition of the latent class variable leads to a log-likelihood function that cannot be maximized explicitly. Instead, the EM algorithm provides an approximate numerical solution that does not require the computation of derivatives by replacing the maximization of likelihood by the maximization of *expected* likelihood, conditional on observed data and an evolving estimate of the underlying parameters.

The EM algorithm can be described as an unsupervised learning algorithm. The basic idea is that many types of data contain unidentified classes that are approximately normally distributed. The EM-algorithm provides a method that can be used to identify these classes by assuming their existence, and then identifying (in a statistical sense) a set of distributions that are most consistent with the observed data. Here we only briefly describe the basic algorithm. Additional details can be found in the appendix.

Suppose that a statistical model of interest contains a set of observed variables $\mathbf{x}$, a set of latent variables $\mathbf{z}$, and a set of underlying parameters Θ, the value of which are unknown and must be estimated. The EM algorithm works as follows. First, an initial (coarse) estimate of Θ is made and denoted by $\Theta^{(0)}$. Second, the conditional expectation of the log-likelihood function $L(\Theta; \mathbf{x}, \mathbf{z})$ is computed, treating $\mathbf{z}$ as a random variable with a density conditional on both $\Theta^{(0)}$ (or $\Theta^{(k)}$ in the k-th step) and the set of observations $\mathbf{x}$. Third, $\Theta^{(1)}$ ($\Theta^{(k+1)}$ in the k-th iteration) is found so as to maximize the conditional expectation in the second step with respect to Θ. Steps two and three are then repeated until a convergence criteria is reached, at which point $\Theta^{(k)}$ is deemed sufficiently close to the maximizer of the joint likelihood function. The resulting estimate is therefore an approximate maximum likelihood (ML) estimate.

The mapping between this general description and the present application is as follows. First, the observed data $\mathbf{x}$ is the collection of mutual fund performances. Second, the latent variable $\mathbf{z}$ is the sequence of populations from which each alpha is drawn, i.e., the unknown ability of each fund. Third, the set of model parameters (Θ) consists of the means, standard deviations, and weights assigned to each sub-population

$$\Theta = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \\ \sigma_1 & \sigma_2 & \sigma_3 \\ \rho_1 & \rho_2 & \rho_3 \end{pmatrix}. \tag{1}$$

A convenient feature of the mixture of normals model is that the equations of the EM algorithm can be written down explicitly.[1] Moreover, a by-product of the estimation procedure is a set of point estimates of the conditional probabilities that observation i comes from population j

$$p^i_j = f_{y_i|x_i,\Theta}(y_i = j), \tag{2}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. (These probabilities are in fact updated and used as inputs in each step of the EM algorithm in order to compute the conditional expectation). These estimates provide a simple classification scheme: each observation is identified with the maximum of the final estimated conditional probabilities evaluated at $\Theta = \hat{\Theta}$. Alternative classification methods might require that p^i_j exceed a specified threshold $\tau \in [0, 1]$.

Maximum likelihood (ML) estimates possess a number of desirable properties under general conditions.[2] First, ML estimates are consistent, which means they are "arbitrary close"

[1] In more general settings the EM algorithm may involve the numerical evaluation of expectations, which is more computationally intensive. See, for example, [Robert and Casella, 2010].

[2] These properties can be found in any standard econometrics textbook, such as [Hayashi, 2011].

to the true parameter values given a sufficient sample size. Second, ML estimates are asymptotically normal with a theoretically known asymptotic covariance matrix. Third, ML estimates are asymptotically efficient, meaning that Θ^*_{mle} minimizes the mean squared error amongst the set of consistent estimators (i.e., the Cramer-Rao bound is achieved). But of course these results are: (1) asymptotic, and therefore approximations with no general advice regarding accuracy for a given sample size, (2) based on strong assumptions regarding the relevant data-generating process, and (3) in the case of approximate ML estimates, sensitive to the choice of initialization. In our results below we conduct a simple simulation exercise that investigates the finite sample properties of our estimates.

3 Data

We use the CRSP Survivor-Bias-Free Mutual Fund Database to identify open-end, diversified, actively managed mutual funds in existence between 1975 and 2014. We exclude funds identified as international, balanced, sector, bond, money market or index funds. From this collection of funds we identify those with multiple share classes and aggregate their observations, with combined fund returns calculated as monthly TNA-weighted returns. We further restrict our sample by only including funds with at least sixty months of observations, which we do not require to be contiguous. We combine the resulting monthly returns data with Fama/French research returns data obtained via Wharton Research Data Services.

We use the returns data to construct a collection of regression intercept estimates (i.e. alphas) and corresponding t-statistics. To generate these values we use the four-factor model studied by [Carhart, 1997]:

$$r_{i,t} = \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + \epsilon_{i,t}, \tag{3}$$

where $r_{smb,t}$, $r_{hml,t}$, and $r_{mom,t}$ are the month t excess returns of the factor-mimicking portfolios capturing size, value, and momentum premiums. We use t-statistics as our basic measure of fund skill because they account for differences in the precision of estimates and are asymptotically Gaussian. Computing the t-statistics requires a choice of standard error computation. We compute the standard errors of all estimates according to [Newey and West, 1994], which adjusts for both heteroskedasticity and correlation across error terms.

Our returns data and the distribution of t-statistics are described in Table 1. Our data contains t-statistic observations from 2,831 funds. In agreement with most prior studies, the mean alpha of an equal-weighted portfolio of the funds over the full sample period is negative and the model R^2 is close to 1 (e.g. [Barras et al., 2010], [Carhart, 1997]).

4 Results

We begin by applying the EM algorithm to the full data set of mutual fund returns. Next we investigate the possibility of skill existing only over short time periods. We then analyze the distribution of fund performance over time, and check for the existence of persistent skill in the following. We then validate our results through a simple simulation exercise.

Period	$\hat{\alpha}$	$\hat{b}_{mkt}$	$\hat{b}_{smb}$	$\hat{b}_{hml}$	$\hat{b}_{umd}$	R^2
1975-2015	-0.522	0.991	0.233	-0.003	0.020	0.979
	(0.481)	(0.011)	(0.031)	(0.029)	(0.017)	
1975-1980	-0.556	1.025	0.262	-0.071	0.170	0.983
	(1.170)	(0.022)	(0.047)	(0.021)	(0.035)	
1980-1985	0.562	0.902	0.339	-0.183	0.030	0.985
	(0.711)	(0.018)	(0.030)	(0.029)	(0.029)	
1985-1990	1.194	0.907	0.275	-0.219	0.030	0.995
	(0.575)	(0.017)	(0.020)	(0.038)	(0.021)	
1990-1995	0.141	0.970	0.266	-0.092	0.042	0.993
	(0.492)	(0.012)	(0.014)	(0.017)	(0.012)	
1995-2000	-1.416	0.965	0.282	-0.003	0.001	0.989
	(0.808)	(0.014)	(0.021)	(0.028)	(0.020)	
2000-2005	-0.823	1.017	0.193	0.128	0.014	0.979
	(1.308)	(0.034)	(0.023)	(0.024)	(0.024)	
2005-2010	-0.228	1.048	0.216	-0.086	0.008	0.989
	(0.789)	(0.035)	(0.035)	(0.041)	(0.017)	
2010-2015	-1.820	0.984	0.239	-0.023	-0.027	0.993
	(0.628)	(0.017)	(0.014)	(0.027)	(0.013)	
Min.	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	Max.	$\bar{\alpha}$	σ
-4.223	-1.369	-0.554	0.287	3.121	-0.551	1.227

Table 1: Summary statistics. The top panel contains parameter estimates of (3) for an equally-weighted portfolio constructed from the dataset of active-managed equity mutual fund returns. The bottom panel contains descriptive statistics for the distribution of t-statistics obtained by estimating (3) for each fund over the full sample period. Values in parentheses are conventional standard errors.

4.1 Long-run skill

We start by estimating the parameters using t-statistics calculated from the full time series of returns for each fund in our sample, as described in Section 3. The results are shown in Table 2. There are several remarkable features of the estimates. First, the proportion of good and poor performers is roughly the same, with each sub-population making up approximately 8% of the overall population. The estimated size of the good performing class is much larger than the corresponding estimate found in [Barras et al., 2010] (8% v. 2.5%), while the estimated poor performing class is significantly smaller (8% v 23%). Second, the estimates show that the average performing funds have significantly negative t-statistics. This observation calls in to question the assumption that the "typical" funds' performance is benign–investing in these funds is in fact detrimental in the long-run. Lastly, there is a substantial spread between the mean of the poor class and the mean of the good class. While the estimated means are means of t-statistics, and therefore cannot be interpreted as returns, it is nonetheless clear that there is a huge cost to investing in a poor performing fund.

4.2 Short-run skill

Next we consider the evidence for skill over short horizons. Specifically, for each fund we divide the time series of returns in to five-year sub-samples starting in the first year of the overall sample period, 1975. We consider each sub-sample as an individual fund and estimate its t-statistic over each sub-period for which the fund has at least 36 months of observations. This procedure results in 7,806 t-statistic observations. Table 2 shows the maximum-likelihood estimates of the short-run sample. Overall the distribution of short-run skill is remarkably similar to the distribution of long-run skill. The primary difference is that the means of the poor and good performers are slightly more extreme in the short-run than in the long-run. These results match our intuition that extreme short-run performances are moderated over the long-run.

4.3 The evolution of skill

Having seen the evidence in favor of a large population of good performers, we next turn to an investigation of how the distribution of fund skill has evolved over time. To do so we perform the following experiment: for each year we repeat the long-run estimation procedure using the entire time series of returns up to the beginning of that year, starting with 1990 and ending with 2015. This results in times series of parameter estimates $\{\mu_t, \sigma_t, \rho_t\}_{t=1990,\ldots,2015}$.

The results are shown in Figure 3. The time series of estimated means do not demonstrate any strong trends over the sample period. On the contrary, the estimates are remarkably stable, particularly over the last ten years. The time series of estimated weights also fail to show any strong trends, and display little variation over the last ten years.

These results are much different from the time series results in [Barras et al., 2010]. In that paper the authors find a strong downward trend in the size of the skilled population, as

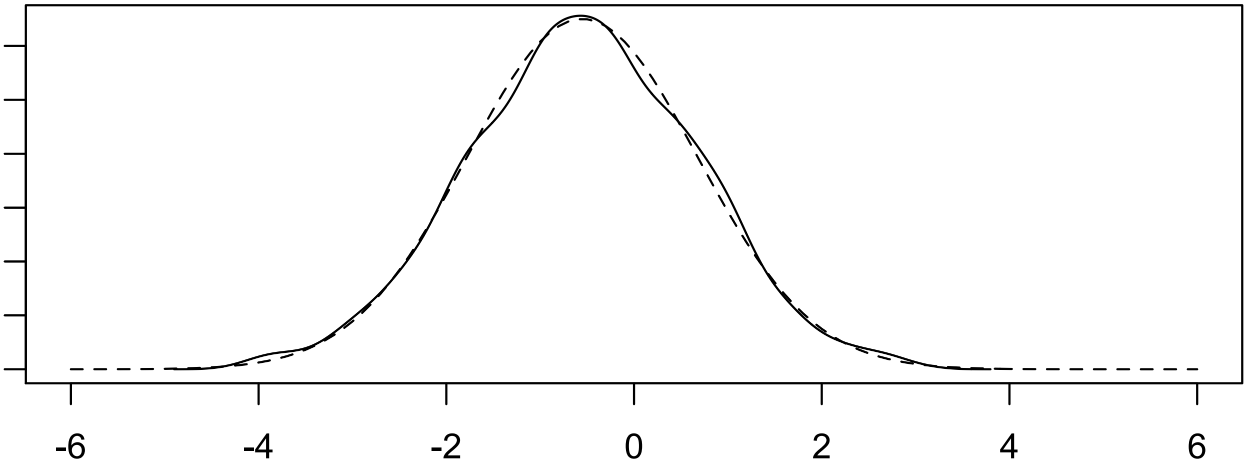


Figure 2: The distribution of t-statistics. The solid curve shows a kernel density estimate of the distribution of t-statistics obtained by estimating (3) for each fund in the data set over the full sample period (1975-2015). The dashed curve shows the best-fitting normal approximation to the same data.

Full Sample: 1975-2015						
	Long-Run			Short-Run		
	Poor	Average	Good	Poor	Average	Good
μ_j	-2.330	-0.550	1.076	-2.640	-0.374	1.650
	(0.304)	(0.071)	(0.123)	(0.092)	(0.037)	(0.283)
σ_j	0.850	1.048	0.852	1.025	1.120	1.136
	(0.129)	(0.032)	(0.057)	(0.038)	(0.019)	(0.120)
π_j	0.078	0.838	0.084	0.091	0.839	0.071
	(0.017)	(0.014)	(0.025)	(0.009)	(0.006)	(0.010)

Table 2: Mixture-model parameter estimates. This table shows the results of our estimation procedure for the sample of actively managed equity mutual funds. The t-statistics used for the *Long-Run* results were obtained by estimating (3) using each funds full monthly returns history ($N = 2,831$). The t-statistics used for the *Short-Run* results were obtained by first dividing each fund's returns history into five-year sub-samples, and then estimating (3) using each sub-sample ($N = 7,806$). Values in parentheses are bootstrapped standard errors.

well as a corresponding increase in the population of unskilled funds. While a reduction in the size of the skilled fund population is consistent with the theoretical predictions found in [Berk and Green, 2004] and [Pastor and Stambaugh, 2012], a dramatic increase in the size of the unskilled population requires further explanation.

4.4 The persistence of skill

Our results thus far show that the distribution of skill is consistent with the existence of sizable sub-populations of good, poor, and average performing funds, and that the properties of these classes are stable over time. It is possible, however, that the constituents of the sub-populations are not stable, making it difficult to capitalize on the performance of the best funds. In order to address this issue we use a classification scheme to construct portfolios of good performing funds and evaluate their out-of-sample performance.

Starting January 1, 1980, each year we form estimates of the skill distribution parameters using the preceding five years of returns. We include all funds that have at least 36 months of returns recorded over this period. Using these estimates we calculate the conditional probability that each fund is a good performer given its estimated t-statistic. We then construct five equal-weighted portfolios of good performers based on fixed probability thresholds $\tau = 0.50, 0.60, \ldots, 0.90$. The classification rule is simple: $P(z_i = j; \hat{\Theta}) > \tau \Rightarrow \hat{z}_i = j$. If no funds are classified as good for a particular threshold level, then the returns over the holding period for that portfolio are the monthly market returns ($\alpha = 0$). Otherwise the portfolio is held for one year, after which the classification procedure is repeated. For comparison we form corresponding portfolios of poor performing funds, as well as a portfolio consisting of all funds that are not classified as either good or poor (which we label "average").

Table 4 describes the composition of the resulting portfolios in terms of the proportion of funds assigned to each portfolio. The proportion of funds in each portfolio is generally

$P(z_i \in \mathit{Good} \mid \hat{\Theta}) \geq \tau$	fund i assigned to population $\mathit{Good}(\tau)$
$P(z_i \in \mathit{Bad} \mid \hat{\Theta}) \geq \tau$	fund i assigned to population $\mathit{Poor}(\tau)$
$P(z_i \in \mathit{Bad}$ or $\mathit{Good} \mid \hat{\Theta}) < 0.50$	fund i assigned to *Average*.

Table 3: Classification scheme.

		Proportion				
Pop. (τ)	$\bar{\pi}_t$	$= 0\%$	$0 - 6\%$	$6 - 12\%$	$12 - 24\%$	$> 24\%$
Poor (0.90)	0.026	13	21	0	0	1
Poor (0.80)	0.037	6	26	2	0	1
Poor (0.70)	0.050	2	27	5	0	1
Poor (0.60)	0.063	0	28	4	2	1
Poor (0.50)	0.079	0	21	10	2	2
Average	0.838	0	0	0	2	33
Good (0.50)	0.083	0	23	8	2	2
Good (0.60)	0.070	0	27	5	1	2
Good (0.70)	0.059	3	26	3	1	2
Good (0.80)	0.047	8	23	1	1	2
Good (0.90)	0.035	14	18	1	1	1

Table 4: Temporal distribution of weights. This table describes the distribution of the weights of the classification scheme described in Table 3. $\bar{\pi}_t$ is the time series average of the proportion of funds assigned to the corresponding population with probability τ. The remaining values are the number of years for which the estimated proportion of funds in the corresponding population ($\hat{\pi}_t$) falls in the given interval. For example, 33 out of 35 years the proportion of funds classified as *Average* exceeds 24%.

smaller than the estimated proportion of funds belonging to the corresponding class due to the threshold criteria. In fact, in many years no funds are assigned to the portfolios constructed using the highest threshold: 13 and 14 out of 35 years for the good and poor classes, respectively.

Table 5 shows the turnover of each portfolio. For each portfolio we calculate the probability that a fund assigned to that portfolio is re-assigned to that portfolio in future years. The results show that there is mild persistence over short time horizons of up to two years, and that the level of persistence decreases as the threshold increases. For example, at the 80% threshold level we find that there is a 20% chance that a fund is re-assigned to either of the good or poor classes one year after it is assigned to the same portfolio. At the 90% threshold level that probability is more than halved. (In fact, the probability that no fund is assigned to this portfolio is nearly twice as high.)

Table 6 analyzes the performance of each portfolio. We evaluate the performance of each portfolio using the same four-factor model used to calculate individual fund t-statistics. The results show that all of the good portfolios generate positive out-of-sample alphas, although none of these are statistically significant by conventional standards. Contrary to what one might expect, the alphas are not increasing (decreasing) in τ for the good (poor) portfolios.

	Years After Classification				
Pop. (τ)	1	2	3	4	5
Poor (0.90)	0.058	0.059	0.044	0.020	0.013
Poor (0.80)	0.196	0.115	0.060	0.037	0.033
Poor (0.70)	0.265	0.162	0.100	0.068	0.049
Poor (0.60)	0.327	0.203	0.129	0.088	0.056
Poor (0.50)	0.368	0.244	0.154	0.107	0.069
Average	0.899	0.854	0.814	0.783	0.751
Good (0.50)	0.343	0.192	0.100	0.075	0.042
Good (0.60)	0.294	0.161	0.076	0.061	0.035
Good (0.70)	0.239	0.131	0.053	0.054	0.032
Good (0.80)	0.193	0.117	0.037	0.053	0.026
Good (0.90)	0.091	0.101	0.021	0.038	0.011

Table 5: Classification turnover. The values in this table are estimates of the probability that a fund assigned to a given population at time t is assigned to the same population at time $t + s$ for $s = 1, \dots, 5$, based on the classification scheme in Table 3. For each formation year between 1980 and 2010, and each population, we calculate the proportion of funds reassigned to the that population $1, \dots, 5$ years later, and then estimate the probability as the simple average of these proportions across all formation years.

The reasons for this are: (a) the classification is imperfect, and (b) the default passive investment is more likely to be used for higher τ's. We also note that the performances of the poor class is highly statistically significant, as is the performance of the average portfolio. Thus our classification procedure provides a substantial benefit to an investor choosing amongst actively managed funds.

Figure 4 shows the performance of the good and poor portfolios over time. With the time series of monthly portfolio returns we construct time series of portfolio alphas by running four-factor regressions at the beginning of each year, starting in Jan. 1990. The time series for the good performers tells a different story from Figure 3: all of the portfolio alphas decrease steadily, suggesting that the size or ability of the good performing sub-population has shrunk over this part of the sample. The discrepancy may be due to the fact that our earlier results are based on expanding windows, but is also a reflection of the high turnover documented in Table 5.

Care needs to be taken when interpreting the results shown in Figure 4. What the figure shows is that prior to 1990 the average performance of the good portfolios was in fact extremely good. The fact that the time series is decreasing over the sample period shown indicates that the portfolio alphas over this period are at best unimpressive, and perhaps negative. When we compare the time series of the classified portfolios with the performance of the naive portfolio, the fact that the alpha of the naive portfolio decreases more indicates that the performance of that portfolio is particularly bad. The reason for this portfolio's poor performance is that it does not account for changes in the distribution of skill. The classification-based portfolios, on the other hand, automatically account for such changes, and switch from active to passive strategies when justified.

Panel A

Pop. (τ)	$\hat{\alpha}$	p	$\hat{b}_{mkt}$	$\hat{b}_{smb}$	$\hat{b}_{hml}$	$\hat{b}_{umd}$	$\bar{r}_p$	σ_p
Poor (0.90)	-2.004	0.012	1.005	0.104	0.030	0.039	6.388	16.44
Poor (0.80)	-2.198	0.002	1.018	0.157	0.023	0.034	6.310	16.73
Poor (0.70)	-2.745	0.000	0.981	0.132	0.045	0.026	5.458	16.12
Poor (0.60)	-3.028	0.000	0.993	0.146	0.039	0.033	5.322	16.40
Poor (0.50)	-2.765	0.000	0.997	0.157	0.033	0.019	5.509	16.34
Average	-1.089	0.007	0.993	0.213	-0.010	0.016	7.064	16.36
Good (0.50)	0.218	0.669	0.955	0.334	-0.044	0.004	8.044	16.55
Good (0.60)	0.561	0.274	0.943	0.332	-0.041	-0.010	8.200	16.37
Good (0.70)	0.421	0.489	0.956	0.332	-0.053	-0.013	8.094	16.73
Good (0.80)	0.706	0.283	0.956	0.268	-0.028	-0.032	8.234	16.44
Good (0.90)	0.540	0.489	0.953	0.203	-0.001	-0.030	8.062	16.21

Panel B

Pop. (τ)	$\bar{\alpha}_t$	$\sigma(\alpha_t)$	$q_{0.05}$	$q_{0.10}$	$q_{0.50}$	$q_{0.90}$	$q_{0.95}$
Poor (0.90)	-1.898	0.378	-22.449	-14.690	0.000	9.075	14.971
Poor (0.80)	-2.834	0.363	-20.011	-14.888	-0.792	7.792	13.095
Poor (0.70)	-3.812	0.424	-28.143	-17.555	-2.455	8.154	14.476
Poor (0.60)	-4.141	0.439	-29.557	-19.982	-3.251	8.737	15.838
Poor (0.50)	-3.473	0.372	-23.910	-16.374	-2.387	8.744	15.077
Average	-1.049	0.253	-13.075	-10.116	-1.358	7.905	11.265
Good (0.50)	0.276	0.336	-15.559	-11.470	0.481	11.181	16.252
Good (0.60)	0.413	0.339	-15.511	-11.613	0.703	11.810	17.500
Good (0.70)	0.571	0.381	-16.952	-12.667	0.000	13.901	17.737
Good (0.80)	0.523	0.337	-14.667	-11.603	0.000	12.606	17.777
Good (0.90)	0.767	0.350	-17.443	-12.019	0.000	13.997	20.727

Table 6: Out-of-sample performance. This table describes the performance of investment strategies based on the classification scheme in Table 3. At the beginning of each formation year equal-weighted portfolios are constructed based on fund performance over the preceding 60 months, and these portfolios are held for the following 12-month holding period. If no funds are assigned to a portfolio in a particular year, then the monthly returns of that portfolio are replaced by the returns of the Russell 5000 index. If a fund disappears during a holding period, then its weight is reassigned to the remaining funds. Panel A shows the parameter estimates of the four-factor model (3) for each resulting time series of monthly returns, as well as the annualized means ($\bar{r}_p$) and standard deviations (σ_p). $\hat{\alpha}$ is annualized. In Panel B we calculate monthly alphas (α_t) for each class-probability portfolio as $r_s - \hat{\beta}_t^T f_s$, where $\hat{\beta}_t$ is an estimate of the four-factor loading using the 60 monthly returns preceding the formation date t. $\bar{\alpha}_t$ and $\sigma(\alpha_t)$ are the mean and standard deviation of the resulting time series of $alpha_t$. q_p is the p-th quantile of the α_t time series. All values are annualized.

Based on the results in Table 5 we know that the composition of our portfolios changes frequently and dramatically. Instead of measuring performance based by alpha as in Table

6, which assumes that the factor loadings of the portfolios do not vary over time, we look at errors based on continually evolving measurements of factor loadings. Specifically, each year after forming our portfolios we calculate the factor loadings from a four-factor model, $\mathbf{B}_t$, and then compute the monthly alphas for each of the next twelve months as $\alpha_{t+i} = r_{t+i} - \mathbf{B}_t\mathbf{f}_{t+i}$ $(i = 1, 2, \ldots, 12)$. This procedure results in a time series of monthly alphas, which we then analyze. The results of this approach are shown in Table 6. We find that the average annualized monthly alpha of portfolios of poor and average performers is substantially lower than that of portfolios of good performers, and that all portfolios of good performers have positive average alphas. The basic message of these results agrees with Table 6: despite high turnover, the average performance of classified portfolios is consistent with their classification and the estimates in Table 2.

4.5 Estimate validation

Our estimation method produces numerical approximations to maximum likelihood estimates. Theoretically, maximum likelihood estimates have desirable asymptotic properties, namely they are efficient estimates. The EM algorithm may, however, struggle to correctly learn the parameters of a mixture of normal model if any of the weights are especially small, or if the densities of the individual distributions have a large amount of overlap (i.e., the means of the distributions are "close" relative to their standard deviations), because in these cases the algorithm is more likely to converge to a local maximizer. Therefore in order to validate our estimation procedure we conduct a simple experiment in which we estimate the parameters of mixture of normal distributions from simulated data.

Specifically, for a fixed set of parameters we generate 1000 samples containing 3000 pseudo-random observations each. Using the exact same implementation of the EM algorithm as we used to obtain our previous estimates, we estimate the known parameters from each sample. We are then able to investigate the finite-sample distribution of our estimators using the set of 1000 estimates.

The results of this experiment are shown in Table 7. We conducted the experiment using two choices of parameters: a set based on inferred values found in [Barras et al., 2010], and a set based on our own estimates. The results show that most of the parameters are estimated with no bias. The exceptions are the mean and weight of the good performers under the distribution based on [Barras et al., 2010]. In this case the population of good funds is simply too small to identify–in fact, the proportion estimated by the authors is not statistically different from zero.

These results present a problem: if the mixture of normal model is approximately correct, then our estimated weights should be highly similar to those found in [Barras et al., 2010]. In Table 2, Panel B we show that, using the same sample period as those authors, this is not the case. A possible explanation for the discrepancy is that the estimation procedure(s) are sensitive to the treatment of the average sub-population's mean. We may find more similar results if we restrict the mean of this class to zero.

Panel A						
	Estimate			Bias		
	Poor	Average	Good	Poor	Average	Good
μ_j	-2.5109	-0.0213	2.2712	-0.0109	-0.0213	-0.7288
σ_j	0.9873	0.9877	1.2333	-0.0127	-0.0123	0.2333
π_j	0.2293	0.7332	0.0375	-0.0007	-0.0168	0.0175

Panel B						
	Estimate			Bias		
	Poor	Average	Good	Poor	Average	Good
μ_j	-2.3764	-0.5435	1.1492	-0.0264	0.0065	0.0692
σ_j	0.8185	1.0346	0.7995	-0.0315	-0.0154	-0.0505
π_j	0.0844	0.8363	0.0793	0.0044	-0.0037	-0.0007

Table 7: Simulation results. In this table we analyze the reliability of our estimates through a simulation exercise. We use the same estimation procedure used in Table 2 to estimate the parameters of a mixture model from data that is randomly generated from a mixture model with known parameter values. For each choice of parameter values we generate 1000 random samples containing 3000 observations each. We report the mean of the 1000 estimates as well as the bias (the mean estimate minus the true parameter value). In Panel A the true distribution is

$$\Theta = \begin{pmatrix} -2.50 & 0.00 & 3.00 \\ 1.00 & 1.00 & 1.00 \\ 0.23 & 0.75 & 0.02 \end{pmatrix}.$$

This distribution is approximately the distribution of t-statistics suggested in [Barras et al., 2010]. In Panel B the true distribution is

$$\Theta = \begin{pmatrix} -2.35 & -0.55 & 1.08 \\ 0.85 & 1.05 & 0.85 \\ 0.08 & 0.84 & 0.08 \end{pmatrix},$$

which is approximately the distribution we estimated in Table 2.

5 Conclusion

This paper has provided a novel assessment of the distribution of skill within the universe of actively managed equity mutual funds. We've specifically addressed two issues of particular academic and practical importance: "Are there *any* funds worth investing in?" And if so, "Can we reliably identify such funds?" Our approach to answering these questions is unique in that it employs a machine learning algorithm to learn about the distribution of fund performance. The method we employ possess several advantages over an alternative methodology proposed in previous, related work. We are able to directly estimate features of the distribution of skill other than the size of potentially skilled or unskilled subpopulations; the method is simple to implement; it leads to an intuitive classification system as well as estimates of the false discovery rate associated with multiple hypothesis tests.

We come up with several interesting findings. First, we find that over both long and short time horizons the distribution of fund performance is consistent with a much larger population skilled funds than previously thought. Unlike previous studies, we impose no restrictions on the mean of the average fund in our sample. We find that while the population of poor performers is no larger than the population of good performers, the average performance of the majority population of funds does not generate excess returns, net of fees.

While the data is consistent with the existence of several distinct classes of skill, we find that it is difficult to reliably classify funds from the information "learned" through our estimation procedure. Indeed, the performance of our investment strategy designed to isolate the class of skilled funds only marginally outperforms a naive investment in the top decile of funds. The reason is that while the *distribution* of skill is stable over our sample period, the *composition* of the individual classes is not. Thus, we do not find evidence of "skilled funds" so much as we find persistent evidence of funds that are temporarily enjoying success.

What are we to conclude of the active management industry? On the one hand it is reasonable to believe that there is superior performance to be found in the distribution of funds at any particular time. On the other hand, it is difficult to identify the funds that will perform well in the near future because the size and excess return of this population are small. But if an investor is, for some reason, constrained to this asset class, then there are benefits to identifying good performers and avoiding poor and average performers.

In this paper we have attempted in this paper to introduce a technique that is well-known in the machine learning community to a finance audience. We believe that machine learning techniques have a role to play within the sphere of financial research and practice, and we look forward to future applications to the study of mutual funds, and financial research generally, in the future.

References

[Baks et al., 2001] Baks, K. P., Metrick, A., and Wachter, J. (2001). Should investors avoid all actively managed mutual funds? a study in bayesian performance evaluation. *The Journal of Finance*, 56(1):45–86.

[Barras et al., 2010] Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: measuring luck in estimated alphas. *The Journal of Finance*, 65:179–216.

[Berk and Green, 2004] Berk, J. B. and Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112:1269–1295.

[Carhart, 1997] Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52:57–82.

[Chen et al., 2015] Chen, Y., Cliff, M. T., and Zhao, H. (2015). Hedge funds: the good, the bad, and the lucky. Available at SSRN: http://ssrn.com/abstract=1915511.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

[Fama and French, 1993] Fama, E. F. and French, K. R. (1993). Common risk factors in the returns of stocks and bonds. *Journal of Financial Economics*, 33:3–56.

[Gruber, 1996] Gruber, M. J. (1996). Another puzzle: The growth in actively managed mutual funds. *Journal of Finance*, 51:783–810.

[Hayashi, 2011] Hayashi, F. (2011). *Econometrics*. Princeton University Press, Princeton, Princeton.

[Jensen, 1968] Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2):389–416.

[Kosowski et al., 2006] Kosowski, R., Timmermann, A., Wermers, R., and White, H. (2006). Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of Finance*, 61:2551–2595.

[Naim and Gildea, 2012] Naim, I. and Gildea, D. (2012). Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients. In *International Conference on Machine Learning (ICML)*.

[Newey and West, 1994] Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61:631–653.

[Pastor and Stambaugh, 2012] Pastor, L. and Stambaugh, R. F. (2012). On the size of the active management industry. *Journal of Political Economy*, 120:740–781.

[Robert and Casella, 2010] Robert, C. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer, New York.

[Storey, 2002] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, 64:479–498.

Appendix

The Expectation-Maximization Algorithm

Assume $\mathbf{x}$ and $\mathbf{z}$ are data with likelihood $L(\Theta|\mathbf{x},\mathbf{z}) = f(\mathbf{x},\mathbf{z}|\Theta)$, but that only $\mathbf{x}$ is observed, and that $\mathbf{x}$ has a likelihood function $L(\Theta|\mathbf{x}) = g(\mathbf{x}|\Theta) = \int f(\mathbf{x},\mathbf{z}|\Theta)\, d\mathbf{z}$. Then the conditional density of the unobserved data given the observed data is

$$k(\mathbf{z}|\mathbf{x}, \Theta) = f(\mathbf{x}, \mathbf{z}|\Theta)/g(\mathbf{x}|\Theta). \tag{4}$$

Given an initial guess of Θ (say $\Theta^{(0)}$) the EM-algorithm consists of the following two-step iterations: (1) construct the expectation

$$Q(\Theta|\Theta^{(k)}) := \mathbb{E}\left[\log f(\mathbf{x}, \mathbf{z}|\Theta) \mid \mathbf{x}, \Theta^{(k)}\right], \tag{5}$$

and (2) update $\Theta^{(k)}$ by computing

$$\Theta^{(k+1)} := \text{argmax}_{\Theta}\ Q(\Theta|\Theta^{(k)}). \tag{6}$$

Note that $Q(\Theta|\Theta^{(k)})$ is the expectation with respect to $\mathbf{z}$ using the conditional density

$$k(\mathbf{z}|\mathbf{x}, \Theta^{(k)}) = f(\mathbf{x}, \mathbf{z}|\Theta^{(k)})/g(\mathbf{x}|\Theta^{(k)}). \tag{7}$$

Steps (1) and (2) are repeated until a convergence criteria is reached. If K iterations are performed, then the MLE estimate is $\Theta^{(K)}$. The posterior density of z_i is found by plugging $\Theta^{(K)}$ in to (7).

Application to a Mixture-of-Normals Model

A normal mixture model refers to a marginal density function

$$g(x_i|\theta) = \sum_{j=1}^{m} p_j \phi(x_i|\mu_j, \sigma_j^2), \tag{8}$$

where $p_1, \ldots, p_m \geq 0$, $\sum p_j = 1$, and $\phi_j := \phi(x_i|\mu_j, \sigma_j^2)$ is a normal density. Observations of $\mathbf{x}$ from the density (8) are equivalent to observations of x_i and z_i whereby z_i is selected such that

$$P(z_i = j) = p_j, \tag{9}$$

for $j = 1, \ldots, m$, and then x_i is drawn from the density ϕ_j density. The joint density of x and z can therefore be written as

$$\sum_{j=1}^{m} \mathbb{I}(z_i = j) p_j \phi_j(x_i), \tag{10}$$

and the likelihood of the data $\mathbf{x}$ and $\mathbf{z}$ is

$$\prod_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}(z_i = j) p_j \phi_j(x_i). \tag{11}$$

Some calculation leads to the log-likelihood function

$$l(\Theta|\mathbf{x}, \mathbf{z}) = \sum_i \sum_j \mathbb{I}(z_i = j) \left[\log p_j - \log \sigma_j - \frac{1}{2}\left(\frac{x_i - \mu_j}{\sigma_j}\right)^2\right]. \tag{12}$$

The conditional expectation of (12) is

$$Q(\theta|\hat{\theta}^{(k)}) = \sum_i \sum_j \; \hat{p}_{ij}^{(k)} l(\Theta|\mathbf{x}, \mathbf{z}), \tag{13}$$

where

$$\hat{p}_{ij}^{(k)} = \mathbb{E}\left[\mathbb{I}(z_i = j) \mid X = x_i, \theta = \hat{\theta}^{(k)}\right] = \frac{\hat{p}_j^k \phi_j^{(k)}(x_i)}{\sum_l \hat{p}_l^{(k)} \phi_l^{(k)}(x_i)}. \tag{14}$$

The maximizer of (13) is

$$\hat{\mu}_j^{(k+1)} = \frac{\sum_i \hat{p}_{ij}^{(k)} x_i}{\sum_i \hat{p}_{ij}^{(k)}} \tag{15}$$

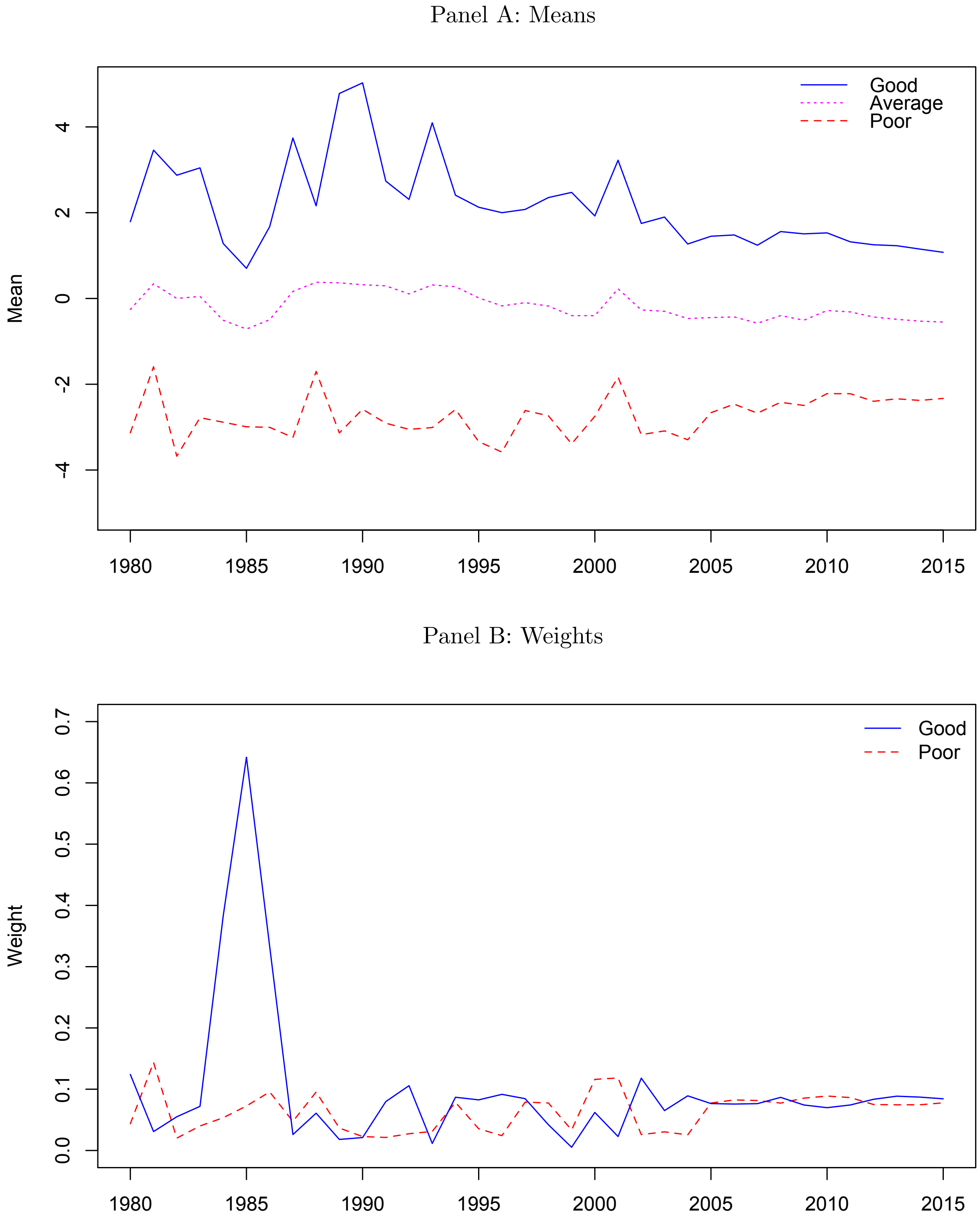
$$\hat{\sigma}_j^{(k+1)} = \sqrt{\frac{\sum_i \hat{p}_{ij}^{(k)} \left(x_i - \hat{\mu}_j^{(k+1)}\right)^2}{\sum_i \hat{p}_{ij}^{(k)}}} \tag{16}$$

$$\hat{p}_j^{(k+1)} = \frac{1}{n} \sum_i \hat{p}_{ij}^{(k)} \tag{17}$$

We allow the algorithm to run until the relative change in the likelihood function is less than $\epsilon = 10^{-6}$. The resulting estimate $\hat{\theta}$ is used to classify each observation by choosing the maximum of

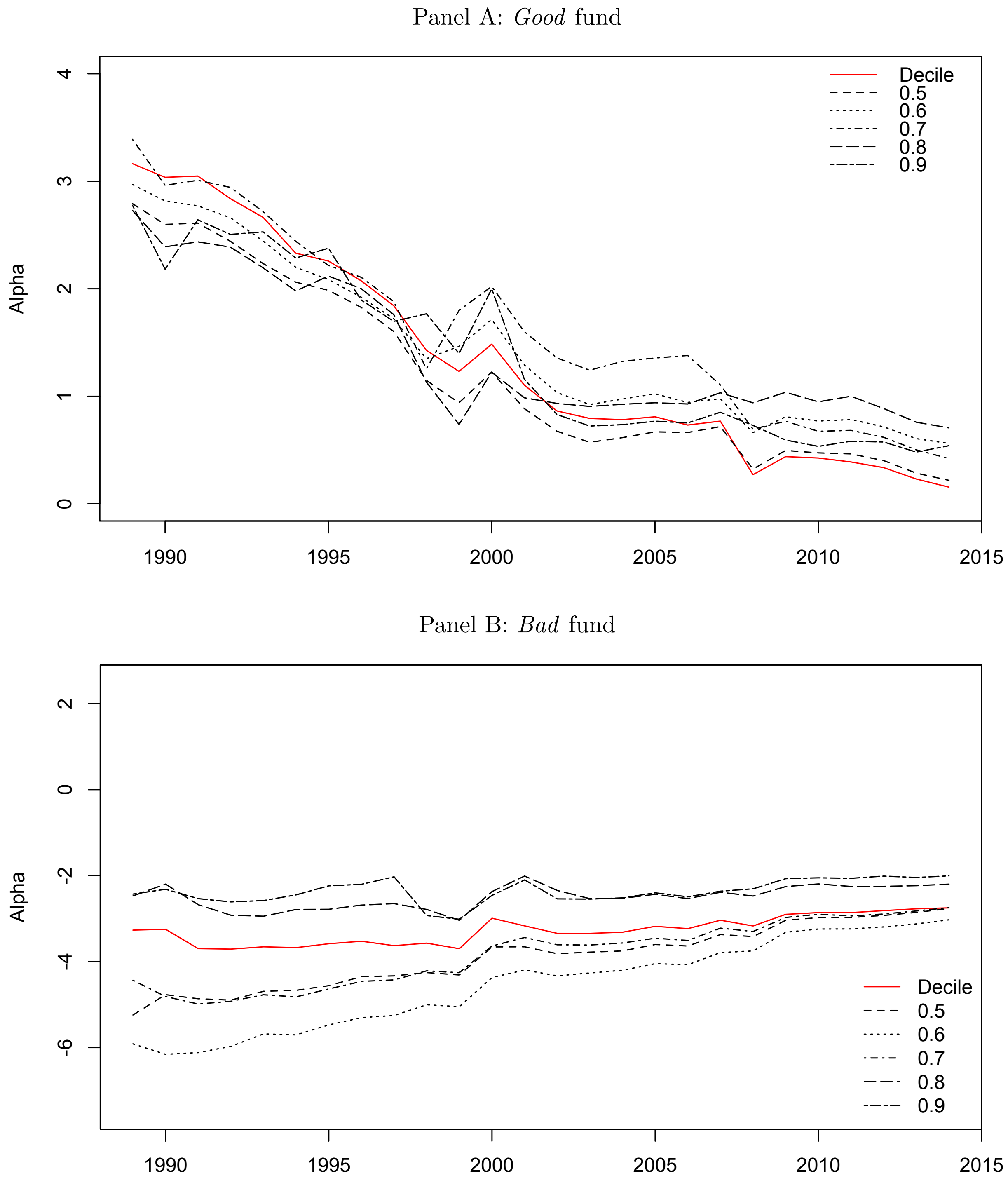
$$\hat{p}_{ij} = \mathbb{E}\left[\mathbb{I}\left(z_i = j\right) | X = x_i, \theta = \hat{\theta}\right] \tag{18}$$

for $j = 1, \ldots, m$.


Panel A: Means

Panel B: Weights

Figure 3: Time series of parameter estimates. We construct time series' of estimates of the parameters in (1) by repeating our estimation procedure each year (starting in 1980) using the full history of monthly returns for each fund available up to that point. The resulting time series of estimated mean t-statistics (μ_i) is shown at top; the time series of estimated weights (ρ_i)is shown at bottom.


Panel A: *Good* fund

Panel B: *Bad* fund

Figure 4: Time series of investment alphas. At the beginning of each year from 1989 to 2014 we estimate a four-factor alpha for each *Good* and *Poor* fund using the entire history of monthly portfolio returns up to that date, as described in Table 6. Panel A (B) compares the performance of the *Good* funds with an alternative investment strategy that invests in the top (bottom) decile of funds, as determined by each fund's four-factor t-statistic over the preceding 60-month period.