On the Size of the Active Management Industry

by*

Ľuboš Pástor

and

Robert F. Stambaugh

August 26, 2010

Abstract

We argue that the popularity of active management is not puzzling despite the industry's poor track record. Our model features decreasing returns to scale: as the industry's size increases, every manager's ability to outperform passive benchmarks declines. We find that the active management industry can remain large even after significantly negative underperformance. Given the observed performance of active mutual funds, investors' proportional allocation to active management should have shrunk only modestly since 1962. We also find investors face endogeneity that limits their learning about returns to scale and allows prolonged departures of the industry's size from its optimal level.

*The University of Chicago Booth School of Business, NBER, and CEPR (Pástor) and the Wharton School, University of Pennsylvania, and NBER (Stambaugh). We are grateful for comments from Andrew Ang, Lord John Eatwell, Gene Fama, Vincent Glode, Rick Green, Ralph Koijen, Kim Min, Dimitris Papanikolaou, Luke Taylor, Rob Vishny, Guofu Zhou, workshop participants at Michigan State University, Ohio State University, University of Chicago, and University of Pennsylvania, as well as participants in the meetings of the NBER Asset Pricing Program, Western Finance Association, European Finance Association, Cambridge/Penn conference, and Q-Group. Support as an Initiative for Global Markets Visiting Fellow (Stambaugh) at the University of Chicago is gratefully acknowledged.

1. Introduction

Active asset management remains popular, even though its track record has long been unimpressive. Consider equity mutual funds, which manage trillions of dollars. Numerous studies report that these funds have provided investors with returns significantly below those on passive benchmarks, on average.[1] While this track record could help explain the growth of index funds, the total size of index funds is still modest compared to that of actively managed funds.[2] Given the negative track record, one might be puzzled by the enormous size of the active management industry.

We argue that the popularity of active management is not puzzling despite its poor track record. Key to this conclusion is to realize that the active management industry exhibits decreasing returns to scale: any fund manager's ability to outperform a passive benchmark declines as the industry's size increases. As more money chases opportunities to outperform, prices are impacted and such opportunities become more elusive. A simple way of modeling returns to scale is as follows:

$$\alpha = a - b\,\frac{S}{W}\,, \tag{1}$$

where α denotes the industry's expected return in excess of passive benchmarks and S/W is the industry's size scaled by investable wealth. Decreasing returns to scale are captured by $b > 0$. If the benchmarks are sufficient for pricing assets in an efficient market, α reflects asset mispricing. In that case, our modeling of decreasing returns to scale is equivalent to assuming that mispricing is reduced as more money seeks to exploit it.

Decreasing returns to scale help us understand the continued popularity of active management. Investors are uncertain about the industry's α, and they learn about it from realized returns. After observing negative benchmark-adjusted returns, investors infer that α is lower than expected, and they reduce their allocation to active management. This reduction in S/W is cushioned by decreasing returns to scale because a lower S/W implies a higher α going forward. Investors infer that α is too low at the current level of S/W, but they know that α will go up after they reduce S/W, so they disinvest less than they would if returns to scale were constant. Under decreasing returns to scale, past underperformance does not imply future underperformance; it only implies that investors should allocate less to active management. After a period of underperformance, the

[1] See Jensen (1968), Malkiel (1995), Gruber (1996), Wermers (2000), Pástor and Stambaugh (2002a), Fama and French (2009), and many others. Fama and French report that, over the past 23 years, an aggregate portfolio of U.S. equity mutual funds underperformed various benchmarks by about 1% per annum.

[2] The Investment Company Institute (2009, p. 20) reports that assets of equity mutual funds total $3.8 trillion at the end of 2008. They also report (ibid., p. 33) that about 87% of those assets are under active management, as opposed to being index funds. Institutions seem more inclined than retail investors to invest passively, but their active allocations are still large, between 47% and 71% of their U.S. equity investments in 2006 (French (2008, Table 3)).

optimal allocation to active management should be smaller than it was at the beginning of the period, but it may remain substantial.

To explore the quantitative implications of the above story, we develop an equilibrium model of active management featuring utility-maximizing investors and fee-maximizing fund managers. We model decreasing returns to scale in a way similar to equation (1), with unknown parameters a and b. After deriving the model's implications for the size of the active management industry, we relate this size to the industry's historical performance.

In our first analysis, we take the familiar perspective on the active-management puzzle, asking whether the industry's size is consistent with its overall level of historical performance. We summarize performance by the t-statistic of the industry's historical alpha, and we construct the posterior distribution for the end-of-sample equilibrium S/W conditional on this t-statistic. We find that the equilibrium S/W can exceed 70% even if the industry's historical alpha is significantly negative ($t = -2$ or less). Intuitively, the t-statistic leaves the researcher quite uncertain about how much historical active returns would have improved had investors allocated less to active management. Given this uncertainty, the confidence region for the equilibrium S/W is wide and it includes allocations that are large. If researchers think that the rational investors in our model could choose a large allocation to active management, it should not puzzle them that actual investors have chosen one.

Whereas our first analysis shows that the active management industry *could* be large despite its poor track record, our second analysis implies that it actually *should* be large. Instead of conditioning on just the overall t-statistic, we now condition year-by-year on the previous history of the industry's returns, as proxied by the returns on the aggregate portfolio of actively managed U.S. equity mutual funds. At the beginning of each year between 1963 and 2006, we solve for the equilibrium S/W that investors in our model would choose if they observed the actual fund returns. We find that this equilibrium allocation drops surprisingly slowly over time—from its assumed value of $S/W = 0.9$ in 1962 to about $S/W = 0.7$ in 2006. This striking result shows that despite substantial past underperformance, active management should remain popular among rational investors facing decreasing returns to scale.

In contrast, active management's popularity would seem quite puzzling under the more traditional assumption of constant returns to scale ($b = 0$ in equation (1)). This assumption is routinely adopted by performance evaluation studies, in which alphas are generally treated as constants, unrelated to the industry's size. We find that under constant returns to scale, the current size of the active management industry should be zero. With $b = 0$, the industry's track record quickly leads investors to perceive $\alpha < 0$ at any S/W, even if their prior beliefs about α are more optimistic

than those leading to the results mentioned above under decreasing returns to scale. With $\alpha < 0$, any positive investment in active management would be undesirable for mean-variance investors; they would instead go short if they could. Our year-by-year analysis with $b = 0$ shows that the equilibrium S/W drops to zero after just seven years and stays there. In other words, if our rational investors thought returns to scale were constant, the active management industry would have disappeared in 1969!

Our proposed reconciliation of the active management industry's large size with its poor track record is the main contribution of this paper. Our second contribution is to show that learning about returns to scale in active management is slow. Investors in our model face endogeneity that limits their learning about a and b in equation (1). As investors update their beliefs about a and b, they adjust S/W. They learn about a and b by observing the industry's returns that follow different allocations. The extent to which they learn is thus endogenous—what they learn affects how much they allocate, but what they allocate affects how much they learn. If S/W ceases to change from one period to the next, learning about a and b essentially stops. Interestingly, we find this is usually the case. The equilibrium S/W converges to the level producing an alpha for the industry that appropriately compensates investors for non-diversifiable risk. Investors eventually learn the alpha at that level of S/W, but they do not accurately learn a and b, even after thousands of years. Convergence of S/W occurs quickly, after just a few years, when b is large. When b is small, though, the industry's size can fluctuate at suboptimal levels for a long time before converging.

Our reliance on decreasing returns to scale in active management owes a debt to the innovative use of this concept by Berk and Green (2004), although our focus and implementation are quite different. Berk and Green assume that an individual fund's returns are decreasing in its own size rather than in the total amount of active management. In their model, as investors update their beliefs about each manager's skill, funds with positive track records attract new money and grow in size, while funds with negative track records experience withdrawals and shrink in size. In reality, actively managed funds have a significantly negative aggregate track record, yet the active management industry remains large. We address this apparent puzzle. Departing from Berk and Green's cross-sectional focus, we analyze the aggregate size of the active management industry.

Another difference from Berk and Green (2004) is our treatment of net fund alphas. Perceived alphas are zero in their model, but they are generally positive in our model, for three reasons. First, alpha reflects compensation for non-benchmark risk that cannot be completely diversified across funds. Such risk is consistent with empirical estimates as well as with the notion that profit opportunities identified by skilled managers are likely to overlap. Second, alpha reflects compensation for uncertainty about the parameters governing the returns to scale in the active management indus-

try. Third, alpha is positive if the number of investors is finite, due to an externality that is inherent to active investing under decreasing returns to scale: each additional investor imposes a negative externality on the existing investors by diluting their returns. When the number of competing investors is large, their lack of coordination drives alpha down, but when their number is small, each investor internalizes a part of the reduction in profits that would result from his own increased investment. We do obtain zero alpha as the limit in the special case in which non-benchmark risk can be completely diversified away (as Berk and Green assume), there is no parameter uncertainty, and the number of investors is infinite.

The equilibrium size of the active management industry depends critically on competition among fund managers. Consider the setting in which there are many investors and many fund managers—the setting on which we mainly focus. The importance of managerial competition is particularly clear in the special case in which investors are risk neutral. The net alpha investors receive in that case is zero whether or not managers compete, but the industry is significantly larger under competition. When managers compete, they become price-takers with respect to their fees, and the industry's equilibrium size produces zero active profit net of those fees. When managers collude, acting monopolistically as one fund, they set the fee rate that produces the fee-maximizing size of the industry in equilibrium. The competitive size exceeds the monopolistic size. In fact, the industry's competitive size is *twice* its monopolistic size if decreasing returns are modeled as in equation (1). If more active management implies less mispricing, then competition among active managers also provides a positive externality to asset markets.

Our study is not alone in trying to explain the puzzling popularity of active management. In our explanation, investors do not expect negative past performance to continue, but in other explanations they do. Gruber (1996) suggests that some "disadvantaged" investors are influenced by advertising and brokers, institutional arrangements, or tax considerations. Glode (2009) presents an explanation in which investors expect negative future performance as a fair tradeoff for countercyclical performance by fund managers. Savov (2009) argues that active funds underperform passive indices but they do not underperform actual index fund investments, because investors buy in and out of index funds at the wrong time. We do not imply that such alternative explanations play no role in resolving the puzzle. We simply suggest that the same job can be accomplished with rational investors who do not expect underperformance going forward.

A number of studies address learning about managerial skill, but none of them consider learning about returns to scale, nor do they analyze the size of the active management industry. Baks, Metrick, and Wachter (2001) examine track records of active mutual funds and find that extremely skeptical prior beliefs about skill would be required to produce zero investment in all funds. They

solve the Bayesian portfolio problem fund by fund, whereas Pástor and Stambaugh (2002b) and Avramov and Wermers (2006) construct optimal portfolios of funds. Other studies that model learning about managerial skill with a focus different from ours include Lynch and Musto (2003), Berk and Green (2004), Huang, Wei, and Yan (2007), and Dangl, Wu, and Zechner (2008).

Our study is also related to that of Garcia and Vanden (2009), who analyze mutual fund formation in a general equilibrium setting with private information. In their model, the size of the mutual fund industry follows from the agents' information acquisition decisions. Asset prices are determined endogenously in their model but not in ours; in that sense, our approach can be described as partial equilibrium, similar to Berk and Green (2004).[3] Recent models of mutual fund formation also include Mamaysky and Spiegel (2002) and Stein (2005). Neither these models nor Garcia and Vanden examine the roles of learning and past data. A number of studies examine equilibrium fee setting by money managers, which occurs in our model as well. Nanda, Narayanan, and Warther (2000) do so in a model in which a fund's return before fees is affected by liquidity costs that increase in fund size. Fee setting is also examined by Chordia (1996) and Das and Sundaram (2002), among others. Finally, whereas our approach is theoretical, Khorana, Servaes, and Tufano (2005) empirically analyze the determinants of the size of the mutual fund industry across countries.

The paper is organized as follows. Section 2 presents our model. After describing the general setting, we first examine the case in which investors are risk neutral. The simple results obtained there for alphas, fees, and industry size clearly reveal the role of competition among managers and investors. The risk-averse case is presented next, followed by the discussion of priors and the updating of beliefs in equilibrium. Section 3 presents the model's quantitative implications for the industry's size given its track record, where we condition first on the overall level of performance and then year-by-year on the previous return history. Section 4 discusses learning about returns to scale. Section 5 relates our model to that of Berk and Green (2004). Section 6 concludes.

2. Model

2.1. Setting

We model two types of agents—fund managers and investors. There are M active fund managers who have the potential ability to identify and exploit opportunities to outperform passive bench-

[3] In addition to Garcia and Vanden (2009), recent examples of studies that analyze the effect of delegated portfolio management on equilibrium asset prices also include Cuoco and Kaniel (2007), Dasgupta, Prat, and Verardo (2008), Guerrieri and Kondor (2008), He and Krishnamurthy (2008), Vayanos and Woolley (2008), and Petajisto (2009).

marks. There are N investors who allocate their wealth across the M active funds as well as the passive benchmarks. The active fund managers' potential outperformance comes at the expense of other investors whose trading decisions are not modeled here.[4]

The rates of return earned by investors in the managers' funds, in excess of the riskless rate, obey the regression model

$$r_F = \underline{\alpha} + Br_B + u, \tag{2}$$

where r_F is the $M \times 1$ vector of excess fund returns, $\underline{\alpha}$ is the $M \times 1$ vector of fund alphas, r_B is a vector of excess returns on passive benchmarks, and u is the $M \times 1$ vector of the residuals. We suppress time subscripts throughout, to simplify notation. Define the benchmark-adjusted returns on the funds as $r \equiv r_F - Br_B$, so that

$$r \;\; = \;\; \underline{\alpha} + u. \tag{3}$$

The elements of the residual vector u have the following factor structure:

$$u_i = x + \epsilon_i, \tag{4}$$

for $i = 1, \ldots, M$, where all ϵ_i's have a mean of zero, variance of σ_ϵ^2, and zero correlation with each other. The common factor x has mean zero and variance σ_x^2. The values of B, σ_x, and σ_ϵ are constants known to both investors and managers.

The factor structure in equation (4) means that the benchmark-adjusted returns of skilled managers are correlated, as long as $\sigma_x > 0$. Skill is the ability to identify opportunities to outperform passive benchmarks, so the same opportunities are likely to be identified by multiple skilled managers. Therefore, multiple managers are likely to hold some of the same positions, resulting in correlated benchmark-adjusted returns.[5] As a result, the risk associated with active investing cannot be fully diversified away by investing in a large number of funds.

The expected benchmark-adjusted dollar profit received in total by fund i's investors and manager is denoted by π_i. Our key assumption is that π_i is decreasing in S/W, where S is the aggregate

[4]The latter investors are required by the fact that alphas (before costs) must aggregate to zero across all investors, an identity referred to as "equilibrium accounting" by Fama and French (2009). These other investors might trade for exogenous "liquidity" reasons, for example, or they could engage in their own active (non-benchmark) investing without employing the M managers. They could also be "misinformed" (Fama and French, 2007) or "irrational" in that they might make systematic mistakes in evaluating the distributions of future payoffs. Such investors might retain a significant fraction of wealth even in the long run, and they can affect asset prices even if their wealth is very small (Kogan, Ross, Wang, and Westerfield, 2006). Good candidates for such investors are individuals who invest in financial markets directly. The proportion of U.S. equity held directly by individuals is substantial: in 1980–2007, this proportion ranged from 22% in 2007 to 48% in 1980 (French, 2008).

[5]This correlation can be amplified if the managers employ leverage because then negative shocks to the commonly employed strategy lead cash-constrained managers to unwind their positions, magnifying the initial shock.

size of the active management industry, and W is the total investable wealth of the N investors. Dividing S by W reflects the notion that the industry's relative (rather than absolute) size is relevant for capturing decreasing returns to scale in active management.[6] In order to obtain closed-form equilibrium results, we assume the functional relation

$$\pi_i = s_i \left(a - b \frac{S}{W} \right), \tag{5}$$

where s_i is the size of manager i's fund, with $S = \sum_{i=1}^{M} s_i$. The parameters a and b in equation (5) are unknown. We denote their first and second conditional moments by

$$\mathrm{E}\left(\begin{bmatrix} a \\ b \end{bmatrix} \mid D \right) = \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} \tag{6}$$

$$\mathrm{Var}\left(\begin{bmatrix} a \\ b \end{bmatrix} \mid D \right) = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}, \tag{7}$$

where D denotes the set of information available to investors.

The parameter a represents the expected return on the initial small fraction of wealth invested in active management, net of proportional costs and managerial compensation in a competitive setting. It seems likely that $a > 0$, although we do not preclude $a < 0$. If no money were invested in active management, no managers would be searching for opportunities to outperform the passive benchmarks, so some opportunities would likely be present. The initial active investment picks low-hanging fruit, so it is likely to have a positive expected benchmark-adjusted return.

The parameter b determines the degree to which the expected benchmark-adjusted return for any manager declines as the fraction of total wealth devoted to active management increases. We allow $b \geq 0$, although it is likely that $b > 0$ due to decreasing returns to scale in the active management industry. As more money chases opportunities to outperform, prices are impacted, and such opportunities become more difficult for any manager to identify. Prices are impacted by these profit-chasing actions of active managers unless markets are perfectly liquid. In that sense, b is related to market liquidity: $b = 0$ in infinitely liquid markets but $b > 0$ otherwise.

We specify the relation (5) exogenously, but decreasing returns to aggregate scale can also arise endogenously in a richer model. In the model of Grossman and Stiglitz (1980), for example, traders can choose to become informed by paying a cost, and the proportion of informed traders is determined in equilibrium. As this proportion rises, expected utility of the informed traders falls relative to that of the uninformed traders, similar in spirit to equation (5).

[6] An alternative way of computing the industry's relative size is S/F, where F denotes the total size of the financial markets. It would seem plausible to assume that π_i is decreasing in S/F. This alternative assumption is equivalent to ours if W grows in fixed proportion to F, which seems like a plausible approximation.

Manager i charges a proportional fee at rate f_i. This is a fee that the fund manager sets while taking into account its effect on the fund's size. The value of f_i, known to investors when making their investment decisions, is chosen by manager i to maximize equilibrium fee revenue,

$$\max_{f_i} f_i s_i. \tag{8}$$

Combining this fee structure with (5), we obtain the following relation for the ith element of $\underline{\alpha}$:

$$\alpha_i = a - b\frac{S}{W} - f_i. \tag{9}$$

The relation between α_i and the amount of active investment is plotted in Figure 1.

Investors are assumed to allocate their wealth across the active funds, the benchmarks, and a riskless asset so as to maximize a single-period mean-variance utility function. We also assume for simplicity that the N investors have identical risk aversion $\gamma \geq 0$ and the same levels of investable wealth. Let δ_j denote the $M \times 1$ vector of the weights that investor j places on the M funds. For each investor j the allocations to the funds solve the problem

$$\max_{\delta_j} \left\{ \delta_j' \mathrm{E}(r|D) - \frac{\gamma}{2} \delta_j' \mathrm{Var}(r|D) \delta_j \right\}, \tag{10}$$

if, as we assume, the allocations to the benchmarks and riskless asset are unrestricted.[7] We impose the restriction that the elements of the $M \times 1$ vector δ_j are non-negative (no shorting of funds).

2.2. Equilibrium under risk neutrality

We first explore the model when $\gamma = 0$. We solve for a symmetric Nash equilibrium among investors, wherein each investor solving (10) takes the optimal decisions of other investors as given. Conditional on the managers' fees, each investor chooses the same vector of allocations, $\delta_j = \delta$, for all $j = 1 \ldots, N$. That solution is then used to compute the equilibrium fees charged by the M managers, who are solving (8). The following proposition gives the equilibrium values of the key quantities in the model.

Proposition 1. *In equilibrium for investors and managers when $\gamma = 0$ and $\tilde{a} > 0$, we have $\mathrm{E}(\alpha_i|D) = \tilde{\alpha}$ and $f_i = f$ for all funds i receiving positive investment. For $M = 1$,*

$$f = \frac{\tilde{a}}{2} \tag{11}$$

[7] As shown in the Appendix, a mean-variance objective for the investor's overall portfolio gives (10) as the relevant problem involving the active funds.

$$\tilde{\alpha} = \left(\frac{1}{2}\right)\frac{\tilde{a}}{N+1} \tag{12}$$

$$\frac{S}{W} = \left(\frac{1}{2}\right)\frac{\tilde{a}}{\tilde{b}}\left(\frac{N}{N+1}\right). \tag{13}$$

For $M > 1$*,*

$$f = 0 \tag{14}$$

$$\tilde{\alpha} = \frac{\tilde{a}}{N+1} \tag{15}$$

$$\frac{S}{W} = \frac{\tilde{a}}{\tilde{b}}\left(\frac{N}{N+1}\right). \tag{16}$$

When $\tilde{a} \leq 0$*, then* $S/W = 0$*. Proof: See Appendix.*

With monopolistic fund management ($M = 1$), the positive fee in (11) is set so that the resulting equilibrium fund size maximizes fee revenue. With competing managers ($M > 1$), the equilibrium fee is $f = 0$. If the fee were instead equal to some positive value, any fund manager setting an infinitesimally lower fee would attract all investment from other funds to that lower-fee fund. Note that f is the portion of a manager's fee that he sets while taking into account its effect on his fund's size. In that sense it is analogous to the part of the price that a supplier sets while taking into account its effect on his sales. Under perfect competition, suppliers and managers are price takers, and such discretionary quantities vanish. That doesn't mean that that suppliers set a zero price or that managers work for nothing. Any competitive proportional fee, which isn't under a manager's discretion, is simply part of a. In other words, a is a rate of return net of proportional costs of producing that return, where the latter costs (not under the manager's discretion) include competitive compensation to the manager and other inputs to producing alpha.

We should note that having f drop to zero as soon as there are even two managers is a result confined to risk neutrality. When M exceeds one but is finite, risk-averse investors do not transfer all of their investments to a manager who sets an infinitesimally lower fee than his competitors, because the other funds still provide diversification that risk-averse investors value. As $M \to \infty$, however, a similar competition argument implies that $f \to 0$. A manager who then sets a fee greater than a common value of $f = 0$ would receive no investment, because investors receive no diversification value from his fund when there are infinitely many other funds.

Competition among the N investors also plays an interesting role. Observe that in both (12) and (15), as $N \to \infty$, $\tilde{\alpha} \to 0$. Alphas become smaller with more investors because each additional investor imposes a negative externality on the existing investors by diluting their returns. Each additional investor does not fully internalize the reduction in alphas caused by the greater amount

invested: his private cost of reducing alphas is less than his private gain from investing. As a result, the greater the number of investors, the larger is the amount of total investment, as is evident in the expressions for S/W in both (13) and (16).

When managers compete, the equilibrium size of the industry in (16) converges to

$$\frac{S}{W} = \frac{\tilde{a}}{\tilde{b}} \tag{17}$$

as $N \to \infty$. With a monopolistic manager, the corresponding limit of (13) is given by

$$\frac{S}{W} = \left(\frac{1}{2}\right)\frac{\tilde{a}}{\tilde{b}}\,, \tag{18}$$

or an industry size only half as large as under competition. The monopolistic size in (18) is also the value that maximizes expected total profit. That is, using equation (5), expected total profit is

$$\Pi = \sum_{i=1}^{M} \pi_i = S\left(\tilde{a} - \tilde{b}\frac{S}{W}\right)\,, \tag{19}$$

which is maximized at the value in (18). In contrast, the competitive industry size in (17) produces zero expected profit in (19). The active management industry in that competitive setting can nevertheless provide a positive externality to asset markets. Suppose the benchmarks are "correct" in an asset-pricing context, in that securities with non-zero alphas with respect to these benchmarks are mispriced. Opportunities to outperform the benchmarks then reflect mispricing. If no money actively chased mispricing ($S = 0$), some mispricing would likely exist. By moving prices toward fair values, the industry provides a positive externality.

In the maximization in (10), we impose the lower bound of zero on the elements of δ_j, but until now we have not imposed any upper bound. A reasonable alternative is to impose the constraint

$$\sum_{i=1}^{M} \delta_{i,j} \leq \delta^*, \tag{20}$$

where $\delta_{i,j}$ denotes the i-th element of δ_j, or the fraction of investor j's wealth invested in fund i. The constraint (20) states that the fraction of each investor's wealth placed in actively managed funds is at most δ^*. When (20) binds, S/W in equation (13) or (16) exceeds δ^*, and the equilibrium value of S/W instead equals δ^*. Also, as in the earlier unconstrained setting, $f = 0$ for $M > 1$: competition among managers drives the discretionary portion of the fee to zero even when the constraint (20) binds. When $M > 1$ and the constraint binds, however, $\tilde{\alpha}$ is a positive value independent of N. In essence, the leverage constraint then prevents investors from increasing the size of the industry to the point at which all profit is eliminated. In contrast, when $M = 1$ and (20) binds, the manager earns a fee greater than the value in (11) and, as in the unconstrained case, $\tilde{\alpha} \to 0$ as $N \to \infty$. The Appendix includes a treatment of the case in which (20) binds.

2.3. Equilibrium under risk aversion

We now turn to a setting with $\gamma > 0$ in the mean-variance objective in (10). To keep the analysis tractable, we confine our attention to the limiting case in which the numbers of managers and investors are both infinite. Relying on the condition $f = 0$ in this competitive setting, we solve for a symmetric Nash equilibrium among investors, each of whom maximizes (10). We obtain an analytic solution for S/W, but the explicit expression—the solution to a cubic equation—is fairly cumbersome. We instead simply present that cubic equation in the following proposition:

Proposition 2. *In equilibrium for an infinite number of investors and managers, if $\tilde{a} > 0$, then S/W is given by the (unique) real positive solution to the equation*

$$0 = \tilde{a} - \frac{S}{W}\left[\tilde{b} + \gamma(\sigma_a^2 + \sigma_x^2)\right] + \left(\frac{S}{W}\right)^2 2\gamma\sigma_{ab} - \left(\frac{S}{W}\right)^3 \gamma\sigma_b^2. \tag{21}$$

If investors also face the constraint in (20) and the solution to (21) exceeds δ^*, then $S/W = \delta^*$. If $\tilde{a} \leq 0$, then $S/W = 0$. Proof: See Appendix.

When the equilibrium value of S/W lies between 0 and 1, it obeys a familiar mean-variance relation. Let r_A denote the benchmark-adjusted return on the aggregate portfolio of all funds. The aggregate analog to the individual investor's problem in (10) is

$$\max_{S/W}\left\{\left(\frac{S}{W}\right)E(r_A|D) - \frac{\gamma}{2}\left(\frac{S}{W}\right)^2 \text{Var}(r_A|D)\right\}. \tag{22}$$

The solution to this problem is given by

$$\frac{S}{W} = \frac{E(r_A|D)}{\gamma \text{Var}(r_A|D)}. \tag{23}$$

To see that (23) characterizes the solution to (21), first note that given the equilibrium value of S/W, the benchmark-adjusted aggregate active fund return is given by

$$\begin{aligned} r_A &= \frac{1}{M}\sum_{i=1}^{M} r_i = \frac{1}{M}\sum_{i=1}^{M}\alpha_i + x + \frac{1}{M}\sum_{i=1}^{M}\epsilon_i \\ &= a - b\frac{S}{W} + x + \frac{1}{M}\sum_{i=1}^{M}\epsilon_i\,, \end{aligned} \tag{24}$$

using equations (3), (4), and (9), and the result that $f = 0$ in equilibrium. Thus, as $M \to \infty$,

$$r_A = a - b\frac{S}{W} + x \tag{25}$$

since the variance of the last term in (24) goes to zero. It follows from (25) that

$$\mathrm{E}(r_A|D) = \tilde{a} - \tilde{b}\frac{S}{W} \tag{26}$$

and

$$\mathrm{Var}(r_A|D) = \sigma_a^2 + \sigma_x^2 - 2\left(\frac{S}{W}\right)\sigma_{ab} + \left(\frac{S}{W}\right)^2\sigma_b^2. \tag{27}$$

Equation (21) can then be rewritten in the image of the mean-variance relation in (23):

$$\frac{S}{W} = \frac{\tilde{a} - \tilde{b}(S/W)}{\gamma\left[\sigma_a^2 + \sigma_x^2 - 2(S/W)\sigma_{ab} + (S/W)^2\sigma_b^2\right]} \tag{28}$$

$$= \frac{\mathrm{E}(r_A|D)}{\gamma\mathrm{Var}(r_A|D)}\,, \tag{29}$$

where the second equality uses (26) and (27).

We can also write equation (25) as $r_A = \alpha + x$, with $\alpha = a - b(S/W)$, so that $\mathrm{Var}(r_A|D) = \sigma_x^2 + \sigma_\alpha^2$, where $\sigma_\alpha^2 = \mathrm{Var}(\alpha|D)$. Equation (29) can then be rewritten as

$$\frac{S}{W} = \frac{\tilde{\alpha}}{\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)} = \frac{\tilde{a} - \tilde{b}(S/W)}{\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)}\,, \tag{30}$$

which gives

$$\frac{S}{W} = \frac{\tilde{a}}{\tilde{b} + \gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)}\,. \tag{31}$$

Note that σ_α^2 depends on S/W, thus requiring the solution to the cubic equation in (21). In the special case where a and b are known, $\sigma_\alpha^2 = 0$ and the right-hand side of (31) yields the solution directly, so solving the cubic equation is then unnecessary.

As before in the risk-neutral solution (17), we see in (31) that greater profitability of the first dollar invested (higher $\tilde{a}$) makes the equilibrium industry size larger, while more strongly decreasing returns to scale (higher $\tilde{b}$) makes the industry smaller. We also see in (31) that greater uncertainty and risk aversion, reflected in the term $\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)$, make the industry smaller. In the rest of this study, we specify risk aversion as $\gamma = 2$ and the volatility of the aggregate active benchmark-adjusted return as $\sigma_x = 0.02$, or 2% per year. The latter value is approximately equal to the annualized residual standard deviation from a regression of the value-weighted average return of all active U.S. equity mutual funds on the three factors of Fama and French (1993), using data for the 1962–2006 period.[8] With these values of γ and σ_x, the value of $\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)$ is often small compared to $\tilde{b}$, especially after some learning about α occurs. As a result, the equilibrium value

[8]The annualized residual standard deviation in that regression, which uses monthly returns, is 1.94%. In a regression of the aggregate active fund return on just the value-weighted market factor, the residual standard deviation is 2.17%. We thank Ken French for providing the series of mutual fund returns, which ends in September 2006.

of S/W is well approximated by (17). The exact values of γ and σ_x do not matter for our study's conclusions; for example, when either value is multiplied by two or divided by two, our results are very similar.

The industry's expected alpha, $\tilde{\alpha} = \mathrm{E}(r_A|D)$, can be obtained by combining equations (31) and (26) to give

$$\tilde{\alpha} \;=\; \tilde{a}\left(\frac{\gamma\left(\sigma_x^2+\sigma_\alpha^2\right)}{\tilde{b}+\gamma\left(\sigma_x^2+\sigma_\alpha^2\right)}\right). \tag{32}$$

When investors are risk averse, we see from (32) that $\tilde{\alpha} > 0$, because investors require compensation for the non-diversifiable risky component x as well as for uncertainty about α.

2.4. Beliefs and Updating

2.4.1. Prior beliefs

We consider a single prior distribution for a but two different prior distributions for b. The first prior for b, or Prior 1, assumes $b = 0$. Prior 1 is a dogmatic belief that returns to scale are constant. The second prior, Prior 2, views b as an unknown quantity satisfying $b \geq 0$. Prior 2 is a belief that returns are decreasing in scale at an uncertain rate. We show below that the two priors lead investors to make very different investment decisions after observing the same evidence.

Both priors can be nested within the joint prior distribution of a and b that is specified below. This joint prior is bivariate normal, truncated to require that $b \geq 0$. That is,

$$\left[\begin{array}{c} a \\ b \end{array}\right] \sim N\left(E_0, V_0\right) I(b \geq 0), \tag{33}$$

where $N(E_0, V_0)$ denotes a bivariate normal distribution with mean E_0 and covariance matrix V_0, and $I(c)$ is an indicator function that equals 1 if condition c is true and 0 otherwise. Denote

$$E_0 = \left[\begin{array}{c} E_0^a \\ E_0^b \end{array}\right], \qquad V_0 = \left[\begin{array}{cc} V_0^{aa} & V_0^{ab} \\ V_0^{ab} & V_0^{bb} \end{array}\right]. \tag{34}$$

Both priors specify $E_0^b = V_0^{ab} = 0$, for simplicity. Prior 1 also specifies $V_0^{bb} = 0$, which implies a degenerate marginal prior distribution for b at $b = 0$. Prior 2 specifies the prior mean of b as $b_0 = 0.2$. Given the properties of the truncated normal distribution, this prior mean implies $V_0^{bb} = 0.063$ and a prior standard deviation for b equal to $\sigma_b^0 = 0.15$. Both marginal prior distributions for b are plotted in the top right panel of Figure 2. Prior 1 appears as a spike at $b = 0$. Prior 2 is the right half of a zero-mean normal distribution truncated below at zero.

Figure 2 also plots the marginal prior distribution for a, in the top left panel. This distribution, which is the same for both Priors 1 and 2, is normal. Its mean and standard deviation, a_0 and σ_a^0, are specified to imply a given prior mean of α at the level of S/W that is optimal under Prior 2. We specify $S/W = 0.9$ as that initial level, so that investors with Prior 2 optimally invest 90% of their wealth in active management before observing any active returns. We choose the prior mean of α equal to $\alpha_0 = 0.1$, or 10% per year, when evaluated at $S/W = 0.9$. Since $\alpha = a - b(S/W)$, the prior mean of a is then equal to $a_0 = \alpha_0 + b_0(S/W) = 0.28$. We choose the prior standard deviation of a such that $S/W = 0.9$ is optimal for investors with Prior 2. Following equation (29), we choose $\sigma_a^0 = \sqrt{\alpha_0/(0.9\gamma) - \sigma_x^2 - (0.9)^2(\sigma_b^0)^2} = 0.19$. Given this large standard deviation, the prior distribution for a is rather disperse, with the 5th percentile at -4% and the 95th percentile at 59% per year. The prior probability that $a < 0$ is 7.2%.

Given the prior distributions for a and b, we can examine the implied priors for α. Since $\alpha = a - b(S/W)$, the prior for α generally depends on S/W. The bottom panels of Figure 2 plot selected percentiles of the prior for α as a function of S/W, which ranges from zero to one. When $b = 0$ (Prior 1, bottom left panel), the distribution of α is invariant to S/W. When $b \geq 0$ (Prior 2, bottom right panel), the distribution of α shifts toward smaller values as S/W increases. The priors for α are fairly noninformative: α might be as large as 60% and as small as -40% per year. Depending on S/W, between 7.2% and 36% of the prior mass of α is below zero.

Importantly, for $S/W = 0$, the prior distribution of α is the same under both priors (because $\alpha = a$ in both cases), but for any $S/W > 0$, α is smaller under Prior 2. In other words, Prior 2 is always more pessimistic about α than Prior 1, at any positive level of S/W. In fact, Prior 1 is so optimistic that in the absence of an upper bound on S/W, investors with that prior would allocate 378% of their wealth to active management before observing any data, as compared to the corresponding allocation of 90% for Prior 2. Despite this prior handicap, investors with Prior 2 generally want to invest more in active management than investors with Prior 1 after observing a negative track record, as we show in Section 3. The reason is that the two priors are updated very differently after observing the same evidence. This updating is described in the following section.

2.4.2. Updating beliefs and equilibrium allocations

Investors update their beliefs in a Bayesian fashion. After each return realization, they compute new posterior moments of a and b, which then yield the updated equilibrium allocation that solves (21). Under Prior 1, where $b = 0$, only the beliefs about a are updated, following the standard result for updating the mean of a normal distribution. Given a history of returns, $y_t = [r_{A,1} \dots r_{A,t}]'$ with

sample average $\bar{r}_{A,t}$, the posterior moments of a (and α) are given by

$$\tilde{\alpha} \;=\; \tilde{a} \;=\; \left(\frac{1}{V_0^{aa}} + \frac{t}{\sigma_x^2}\right)^{-1} \left(\frac{E_0^a}{V_0^{aa}} + \frac{t\bar{r}_{A,t}}{\sigma_x^2}\right). \tag{35}$$

$$\sigma_\alpha^2 \;=\; \sigma_a^2 \;=\; \left(\frac{1}{V_0^{aa}} + \frac{t}{\sigma_x^2}\right)^{-1}. \tag{36}$$

When $b = 0$, the cubic equation in (21) simplifies to a linear equation. In this case, σ_α^2 does not depend on S/W, so the new allocation $(S/W)_{t+1}$ is then given directly by the first equality in (30). The active-management allocation problem in this case is essentially equivalent to the setting in Treynor and Black (1973), but with the addition of parameter uncertainty.

Under Prior 2, investors must infer the coefficients in a time-series regression of returns on the equilibrium allocations. After observing $r_{A,t}$, which is the return following investors' equilibrium allocation $(S/W)_t$, the available data in D consist of y_t and $z_t = [(S/W)_1 \ldots (S/W)_t]'$. In a regression of y_t on $-z_t$ and a constant, the intercept is a and the slope is b (see equation (25)). Recall that investors' prior beliefs for a and b are given by the bivariate truncated normal distribution in equation (33), whose non-truncated moments are E_0 and V_0. In year t, those moments are updated by using standard Bayesian results for the multiple regression model,

$$V \;=\; \left(V_0^{-1} + \frac{1}{\sigma_x^2}\left(Z_t' Z_t\right)\right)^{-1} \tag{37}$$

$$E \;=\; V^{-1}\left(V_0^{-1}E_0 + \frac{1}{\sigma_x^2} Z_t' y_t\right), \tag{38}$$

where $Z_t = [\ \iota_t \quad -z_t\ ]$. The posterior distribution of a and b is bivariate truncated normal as in equation (33), except that E_0 and V_0 are replaced by E and V from equations (37) and (38).[9] Having the updated moments E and V of the non-truncated bivariate normal distribution, we apply the relations in Muthen (1991) to obtain the updated moments of the truncated bivariate normal distribution, defined in equations (6) and (7).[10] Those moments are then used to solve the cubic equation in (21) to obtain $(S/W)_{t+1}$.

[9] In deriving the posterior of a and b from the regression of y_t on $-z_t$, it is useful to note that $(S/W)_t$ is a deterministic function of its initial value and returns prior to time t, so there is no randomness in S/W beyond what is in past returns. The likelihood function is obtained simply by transforming the density of $\{x_s;\ s = 1, \ldots, t\}$ to the density of $\{r_{A,s};\ s = 1, \ldots, t\}$, where the Jacobian of that transformation equals 1. As a result, the likelihood function is identical to what would arise if the observations of S/W were treated as nonstochastic.

[10] Earlier results for such moments appear in Rosenbaum (1961), but the published article contains some errors in signs that we verified through simulation.

3. Is the industry's size puzzling given its track record?

In this section, we use our model to ask whether it is puzzling that the active management industry remains large, given its historical performance. We use historical performance in two different ways. First, we take the perspective of a researcher who computes the t-statistic of the industry's historical alpha. Conditioning on this measure considers the active-management puzzle in its traditional context, wherein the overall level of poor performance for the industry, summarized here by the t-statistic, seems at odds with the substantial size the industry still enjoys. Second, we go beyond this traditional perspective and condition on the industry's actual year-by-year performance. In each year of the historical sample, we solve for the equilibrium allocation to active management, conditional on the history of active returns and equilibrium allocations.

We use the same priors for a and b as presented earlier, Prior 1 ($b = 0$) and Prior 2 ($b \geq 0$). Allowing decreasing returns to scale, as does Prior 2, plays a crucial role in judging the industry's continued popularity, whether we condition on just the t-statistic or on the full year-by-year return series. Of particular interest is S/W at the end of the sample period, which corresponds to the allocation to active management at the present time. When conditioning on just the t-statistic, that equilibrium value of S/W is described by a posterior distribution. We show below that this distribution includes substantial allocations, in excess of 0.7, even when the t-statistic equals -2. Even more striking are the results obtained by conditioning on the entire year-by-year series of actual returns on actively managed U.S. mutual funds. In that case, the equilibrium allocation is given by a single value, which is equal to about 0.7. In other words, not only do we find that the industry *could* still be relatively large given its overall level of negative performance, we find that it *should* still be large given the evolution of that track record. This conclusion critically depends on decreasing returns to scale. If we use the constant-returns-to-scale Prior 1 instead, the current equilibrium size of the industry is zero under either form of conditioning, even though Prior 1 is more optimistic than Prior 2 about active management's α (as discussed earlier).

3.1. Conditioning on the overall level of performance (t-statistic)

We first take the perspective of a researcher who uses the posterior distribution of the current equilibrium S/W, conditional on an overall summary measure of the industry's track record, to judge the reasonableness of the current actual S/W. The researcher knows that the latter quantity is substantial, but he does not observe it precisely. Measuring S/W is difficult from the researcher's perspective, especially because W is difficult to measure. First, W includes cash. Recall that W is allocated across active funds, passive benchmarks, and cash (the riskless asset). The investors'

cash holdings are difficult to pin down. Second, W is only a subset of total wealth; it is the wealth of our N investors. It seems difficult to empirically separate the wealth of these investors from the wealth of the other unmodeled investors discussed at the beginning of Section 2. We do assume that the researcher's prior beliefs about a and b are the same as those held by our investors.

In computing the posterior distribution for the equilibrium S/W conditional on the overall track record, we characterize the track record by the t-statistic of the industry's historical alpha. This historical alpha, or $\hat{\alpha}$, is simply the sample average benchmark-adjusted return. Its t-statistic is computed as $t = \hat{\alpha}\sqrt{T}/\sigma_x$ for $T = 50$ years.[11] For each prior, the posterior distribution of $(S/W)_T$ in year $T = 50$ is obtained by simulating 300,000 samples. To simulate a sample, a and b are drawn from the prior, and then in each year $t = 1, \ldots, T$ a return is drawn following equation (25) as $r_{A,t} = a - b(S/W)_t + x_t$, where $x_t \sim N(0, \sigma_x^2)$. The new equilibrium allocation $(S/W)_{t+1}$ is then computed using the updating procedure described in Section 2. Note that under Prior 2, $(S/W)_{t+1}$ affects $r_{A,t+1}$: the more investors allocate to active management, the lower is their subsequent return. In contrast, there is no such relation under Prior 1. The posterior distribution for $(S/W)_T$ conditional on a given value t_0 of the t-statistic is constructed as the distribution of the $(S/W)_T$ values in all simulated samples producing t-statistics within a small neighborhood of t_0. Figure 3 plots the resulting posterior distributions for t_0 ranging from -4 to 4.

Panel A of Figure 3 displays the posterior distribution of S/W obtained under Prior 1 ($b = 0$), according to which there are constant returns to scale. The posterior distribution collapses to a single value because the t-statistic is a sufficient statistic for S/W in this case. The optimal allocation is a steep linear function of past performance as long as that performance is mildly positive (t-statistics between 0 and 0.25). If past performance is more positive ($t > 0.25$), the optimal allocation is $S/W = 1$. If past performance is negative, we obtain the other corner solution, $S/W = 0$. The cutoff value of the t-statistic that produces $S/W = 0$ is just below zero. It is not exactly zero because the prior for a is slightly informative (see Figure 2), but it is very close to zero. So it is a reasonable approximation to state that investors observing negative past performance optimally choose to invest nothing in active management. This implication of $b = 0$ does not seem to match the reality, in which the active management industry continues to attract substantial investment despite having delivered negative performance relative to passive indices.

The puzzling coexistence of negative past performance and substantial investment is easier to understand when there are decreasing returns to scale. Panel B of Figure 3 plots the posterior distribution of S/W conditional on the t-statistic under Prior 2 ($b \geq 0$). Unlike in Panel A, the t-statistic is no longer a sufficient statistic for S/W. Panel B shows that S/W increases with past

[11] The results for other values of T, such as 20 or 30 years, are very similar.

performance, though not as steeply as in Panel A. When the historical alpha is zero ($t = 0$), the middle 90% of the distribution of S/W (between the 5th and 95th percentiles) lies in the wide range between 0.26 and 0.97. When the historical t-statistic is $t = -2$, indicating statistically significant underperformance, the median S/W is 0.27 and the middle 90% of the distribution ranges from 0.02 to 0.71. Note that $S/W < 0.02$ is as unlikely as $S/W > 0.71$: observing very little investment in active management would be equally puzzling as observing too much investment. Even when the t-statistic is $t = -3$, which is more negative than the observed evidence for mutual funds, the median S/W is 0.13 and the 95th percentile is 0.43. Panel B clearly shows that when $b \geq 0$, substantial investment in active management can be optimal even when past performance is significantly negative.

Investors are willing to invest despite poor past performance because past underperformance does not imply future underperformance. Under decreasing returns to scale, the expected return in any given period is conditional on the investment level S/W in that period. Historical benchmark-adjusted returns are earned at various investment levels, which can be quite different from the current investment level. After a period of underperformance, investors reduce their investment until their expected return going forward converges to the positive equilibrium level of α in equation (32). If past performance is sufficiently poor, investors will choose to invest nothing in active management; this happens if investors infer that a is nonpositive, in which case α cannot be positive either. Such an event occurs with only 13% probability even when $t = -3$, and active management does not seem to have underperformed quite that badly. For the 1962–2006 period, the regression (mentioned earlier) of the value-weighted active U.S. equity fund excess return on the three Fama-French factors produces $t = -1.7$, while a regression on just the market factor produces $t = -2.6$. At such levels of underperformance, the optimal investment in active management can be substantial. For example, when $t = -1.5$, the median S/W is 0.37 and the 95th percentile is 0.84, and when $t = -2.5$, the median is 0.19 and the 95th percentile is 0.56.

It is important to note that optimism about active management's abilities is not our story. If the prior for a in Figure 2 strikes one as optimistic, one should recall that the same prior leads investors who set $b = 0$ to invest nothing in active management given its negative track record, even though the implied prior about α is then more optimistic. Our results are instead driven by the prior on b. Further support for this statement is provided by results we obtain for an alternative set of priors that are less optimistic. Specifically, while keeping the same priors for b as before, we modify the prior for a so that the optimal initial value of S/W under Prior 2 is now 0.5 instead of 0.9. This prior assigns a 26% probability to $a < 0$, compared to the 7.2% probability in Figure 2. Naturally, the implied prior for α is then more pessimistic as well, with the median ranging from 0 to 0.2 and the 5th percentile ranging from -0.6 to -0.3 for $0 \leq S/W \leq 1$. Under this alternative specification,

conditional on $t = -2$, the median S/W after 50 years is 0.10 and the 95th percentile is 0.65. These values are smaller than their counterparts in Figure 3, but that is not surprising. Before seeing any data, investors under this alternative prior allocate only half of their investable wealth to active management, so following a negative track record, they generally allocate less than half. Nevertheless, substantial allocations still lie within the posterior distribution. As before, of course, the same track record implies $S/W = 0$ when investors set $b = 0$.

3.2. Conditioning on the year-by-year returns of U.S. mutual funds

We now take a perspective akin to that of the investors in our model. Recall that investors determine $(S/W)_t$ as a function of the previous active returns $\{r_{A,1}, \ldots, r_{A,t-1}\}$ and equilibrium allocations $\{(S/W)_1, \ldots, (S/W)_{t-1}\}$. For each year t from 1963 through 2006, we set $r_{A,t-1}$ equal to the return on the aggregate portfolio of actively managed U.S. equity mutual funds, net of the return attributable to the portfolio's estimated exposures to the three factors of Fama and French (1993). As in the model, each $(S/W)_t$ is then determined by the returns and allocations through period $t - 1$. Priors are again as specified in Figure 2.

Figure 4 displays the resulting path of S/W over the sample period. When $b = 0$, the value of S/W, which starts at the upper limit of 1 given the prior in this case, bounces between 0 and 1 during the first 7 years but then settles at $S/W = 0$ thereafter. We thus see that, when $b = 0$, not only is S/W equal to 0 conditional on a negative overall track record, as in the previous analysis, but S/W converges to 0 after just 7 years of the track record as it actually occurred. In other words, if one believes investors preclude the possibility of decreasing returns to scale, the active management puzzle becomes even deeper than previously recognized.

In sharp contrast to the $b = 0$ case, the prior with $b \geq 0$ yields an equilibrium allocation that drops rather smoothly and modestly over time, from its initial value of $S/W = 0.9$ to a final value of about $S/W = 0.7$. The intuition behind the slow decline in S/W is simple. When investors observe negative benchmark-adjusted returns, they revise downward their beliefs about the active management's α at the current level of S/W. As a result, investors reduce their allocation to active management. However, this reduction in S/W is mitigated by decreasing returns to scale because a lower S/W implies a higher α going forward. Due to decreasing returns, investors disinvest less in response to poor performance than they would if returns to scale were constant.

The plot in Figure 4 seems plausible. As discussed earlier, it is difficult to measure S/W empirically, but even casual observation would surely indicate that it has been substantial over time and remains so today. At the same time, it seems likely that S/W has declined somewhat over time, given the growth of indexing. For equity mutual funds, indexing has grown from its

inception in the 1970's to its recent share of about 13 percent of total assets (Investment Company Institute, 2009). Among institutional investors, the growth of indexing has been greater, to a recent share between 31 and 53 percent, depending on investor classification (French, 2008). Consistent with these facts, Figure 4 shows an allocation that has dropped somewhat but is still substantial. The striking message from Figure 4 is that rather than the industry's size being a puzzle given its track record, the year-by-year track record actually implies that active management should *not* have shrunk dramatically from an initially substantial allocation of investor wealth.

4. Learning About Returns to Scale

In this section, we analyze how investors learn about decreasing returns to scale in active management. We find that this learning is slow, hampered by an interesting endogeneity faced by competing investors who cannot coordinate their investment decisions. Even though investors eventually learn the industry's α, they never accurately learn a and b in equation (1). We also find that when b is large, the equilibrium allocation to active management stabilizes quickly, but when b is small, the industry's size can fluctuate at suboptimal levels for a long time.

As investors learn, their posterior standard deviations of a, b, and α decline through time. For a given prior, the manner in which these posterior standard deviations decline depends on the true values of a and b. The probability distributions of a and b thus give rise to distributions of the posterior standard deviations of a, b, and α. Figure 5 displays the evolution of these distributions over time. Each panel plots selected percentiles of the distribution of the given standard deviation across the 300,000 samples described earlier near the beginning of Section 3.1.

The three left panels of Figure 5 correspond to Prior 1 ($b = 0$). In this case, a and α coincide, which is why the top and bottom left panels of Figure 5 look identical. (The middle left panel looks empty because the posterior standard deviation of b is zero.) The learning process is straightforward. With $b = 0$, the value of a (and α) is simply the unconditional mean return. The posterior mean of a is a weighted average of the historical average return and the prior mean, where the weight on the prior mean quickly diminishes as t increases because the prior for a is fairly noninformative (Figure 2). The posterior standard deviation of a declines at the usual $\sqrt{t}$ rate, regardless of the particular sample realization. Since there is no dispersion in the standard deviations across the simulated samples, the distribution of the standard deviations collapses into a single line. In short, when $b = 0$, learning is simple and well understood. In contrast, learning is much more interesting when $b \geq 0$, as explained next.

The three right panels of Figure 5 represent Prior 2 ($b \geq 0$). Under this prior, the posterior standard deviations of a and b fall sharply in the first few years but then flatten out surprisingly quickly. For the median sample, investors learn much more about a and b in the first two or three years than in the subsequent 50 years! Moreover, even after 50 years, investors remain highly uncertain about a and b: for the median sample, the posterior standard deviations of a and b both exceed 7%. For comparison, the posterior standard deviation of a is 25 times smaller when $b = 0$. The speed of learning about a is clearly very different when $b \geq 0$ than when $b = 0$ (compare the top two panels in Figure 5).

Investors learn differently under the two priors for b because the level and variation in $(S/W)_t$ affect learning when $b \geq 0$ but not when $b = 0$. We discuss this difference in Section 4.1. This difference is absent, however, when S/W is persistently equal to zero. In 6.3% of all samples, $(S/W)_t = 0$ for all t between 3 and 50 years. These are samples in which investors quickly learn that it is optimal for them to invest nothing at all in active management (because they perceive $a < 0$). In these samples, S/W does not affect learning, just like when $b = 0$, so the results for these samples should look the same between 3 and 50 years whether $b \geq 0$ or $b = 0$. Indeed, Figure 5 shows that the 5th percentile of the posterior standard deviation of a in the top right panel ($b \geq 0$) looks the same as in the top left panel ($b = 0$) after year 3. The same 5th percentile also looks very similar to the 5th percentile of the posterior standard deviation of α in the bottom right panel, again because more than 5% of all samples exhibit $S/W = 0$ and hence also $\alpha = a$.

In contrast to the difference in speeds at which investors learn about a under Prior 1 versus Prior 2, investors learn about α at essentially the same rate under both priors. This similarity in learning speeds is evident in a comparison of the bottom two panels of Figure 5. As will be discussed later, the equilibrium value of S/W under Prior 2 ($b \geq 0$) typically does not change much after just a few years. If S/W stops changing, so does α, so investors with Prior 2 are then able to learn α at that stable S/W about as quickly as investors with Prior 1 learn the value of α that they assume to be constant. In the earlier discussion of equation (31), we noted that the term $\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)$ makes a relatively small contribution compared to the other denominator term $\tilde{b}$, especially after some learning about α has occurred. We now see in the bottom right panel of Figure 5 that σ_α is typically about 2%—the same as σ_x—after the first year's return is observed, and then σ_α drops at essentially the same $\sqrt{t}$ rate as it does in the bottom left panel with $b = 0$. Thus, as t increases, the term $\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)$ is soon well approximated by $\gamma\sigma_x^2$, which is generally small compared to $\tilde{b}$ for the values of γ and σ_x^2 considered.

4.1. Endogeneity in Learning

The key message from Figure 5 is that most of the time, learning about a and b essentially stops after just a few years. The reason is the endogeneity in the way investors learn—what they learn affects how much they invest, and how much they invest affects what they learn. If the amount invested stops changing from one period to the next, investors stop learning about returns to scale. Recall that investors essentially run the time-series regression of active returns, $r_{A,t}$, on the equilibrium allocations to active management, $(S/W)_t$. If the right-hand side variable in the regression stops changing, investors stop learning about the true values of the intercept and slope. Indeed, we find that in most cases, $(S/W)_t$ ceases to change much after just a few years.

The fact that the aggregate active allocation $(S/W)_t$ typically ceases to change reflects equilibrium among competitive investors. If investors could instead coordinate, they might well find it useful to continue varying the aggregate active allocation for additional periods, so as to continue learning about a and b. In a multiperiod setting, such investors would trade off near-term optimality of their current allocation against the potential future value of additional learning by experimenting with different allocations. The additional learning could be valuable, for example, if investors could experience a future preference shock making their previous allocation suboptimal. With learning about a and b shut down, investors are uncertain about α at any allocation other than the current one. The prospect of wanting to change their allocation in the future creates an incentive for additional learning about a and b.

To illustrate the endogenous nature of learning in our competitive setting, Figure 6 plots representative examples of learning paths for various random samples. The figure has 12 panels, each of which plots returns $r_{A,t}$ against $(S/W)_t$ for $t = 1, \ldots, 300$ years. The three columns of panels correspond to three different values of b: "low" (5th percentile of the prior distribution, 0.02), "median" (50th percentile, 0.17), and "high" (95th percentile, 0.49). Given the value of b, the value of a is computed so that the "true" value of S/W that would obtain if the true values of a and b were known is either $S/W = 0.5$ (the top six panels) or $S/W = 0.7$ (the bottom six panels). The true value of S/W is given by $a/(b + \gamma\sigma_x^2)$, which is a special case of equation (31) when a and b are known. The (a, b) pair obtained above is then used to generate random samples of active returns, which are used to update Prior 2. All panels of Figure 6 represent examples of learning paths that commonly occur for the given values of a and b. The starting point ($t = 1$) is indicated with a circle; its x coordinate is always $(S/W)_1 = 0.9$.

The intuition for why $(S/W)_t$ tends to stop changing comes across most clearly when b is high. Consider the top right panel of Figure 6. Since the initial allocation exceeds the true value,

that is $(S/W)_1 = 0.9 > 0.5$, investors initially overinvest in active management, so their true expected return is negative (even though they subjectively expect a positive return). The first realized return is about -19%. Upon observing such a large negative return, investors sharply revise their prior beliefs and dramatically cut their allocation, to about $(S/W)_2 = 0.3$. This represents underinvestment relative to the true S/W, so the realized return in the second year tends to be larger than investors expect, about 9%.[12] From this high return, investors infer they should invest more than 0.3. Their investment in year three, $(S/W)_3$, is already close to the true value of 0.5. In all four panels in which b is high, S/W "converges" to its true value after only three or four years, in that only small deviations from the true value appear over the following 300 years.

Why does the equilibrium allocation approach the true S/W so quickly when b is high? The reason is that after two years, investors already have a lot of information about the true S/W, which is equal to $a/(b+\gamma\sigma_x^2)$, as mentioned earlier. When b is high, the true value is approximately equal to a/b.[13] This approximate relation can be visualized in Figure 1. When b is high, the equilibrium true S/W is very close to $\bar{S}/W = a/b$. The true S/W is slightly smaller than $\bar{S}/W$ (and α is slightly positive) because investors demand compensation for nondiversifiable risk (i.e., because $\gamma\sigma_x^2 > 0$). However, since $\gamma\sigma_x^2$ is small compared to b, α is close to zero and $S/W \approx \bar{S}/W$.

To understand why investors know a lot about $\bar{S}/W$ after two years, recall that $\bar{S}/W$ represents the point at which the line in Figure 1 intersects the x axis. After two years, investors observe two datapoints, $((S/W)_1, r_{A,1})$ and $((S/W)_2, r_{A,2})$, which are far from each other, both vertically and horizontally (because investors update their relatively noninformative prior beliefs substantially after the first observation). Fitting a line through these two distant points allows investors to pin down the intersection point $\bar{S}/W$ reasonably well. As a result, approximate convergence to the true S/W tends to occur quickly when b is high.

This logic also helps us understand the L-shaped pattern in the posterior standard deviations of a and b in Figure 5. As noted earlier, a and b are estimated from the regression of $r_{A,t}$ on $(S/W)_t$. This regression can be visualized as fitting a line through the datapoints plotted in Figure 6, a line whose intercept is a and whose slope is $-b$. In the first few years, investors learn a lot about a and b due to substantial initial variation in S/W. Fitting a line through the first two datapoints already substantially reduces the prior uncertainty about the intercept and the slope. This is why the posterior standard deviations of a and b in Figure 5 exhibit a sharp initial drop.

[12] This systematic underinvestment appears from our perspective because we know the true value of S/W. In contrast, there is no underinvestment (or overinvestment) from the perspective of our investors who do not know the true S/W. The investors always invest optimally given their information set.

[13] Our high value of b, the 95th percentile of the prior for b, is equal to 0.49, which far exceeds $\gamma\sigma_x^2 = 0.0008$.

After the first few years, however, S/W exhibits very little variation when b is high, thereby precluding investors from getting much new information about the intercept and slope. Facing the 300-year data pattern in the top right panel of Figure 6, investors fit a line through what are effectively only three datapoints: $((S/W)_1, r_{A,1})$ from year 1, $((S/W)_2, r_{A,2})$ from year 2, and the midpoint of the cluster of points at $S/W \approx 0.5$ from years 3 through 300. Therefore, investors do not know much more about a and b after 300 years compared to what they knew after 3 years. The same logic also applies when b is not high, albeit to a lesser extent. S/W often settles at a given value for a long period of time, thereby slowing down learning about a and b. This is why the posterior standard deviations in Figure 5 decline so slowly after just a few years.

In the preceding discussion of why $(S/W)_t$ converges quickly, we focus on the high value of b. The four middle panels of Figure 6 show examples of learning paths when b is at its prior median. The main difference from the high b case is that S/W typically fluctuates for several decades rather than years before converging. The convergence of S/W generally takes longer when the initial S/W is closer to the true S/W because learning is then slower due to a smaller magnitude of the initial realized return. (In the middle column, compare examples 1 and 2, in which the true $S/W = 0.5$, with examples 3 and 4, in which the true $S/W = 0.7$.)

4.2. Departures from Optimal Industry Size

To further analyze the convergence of S/W, we examine the distribution across the 300,000 samples of the difference between the equilibrium $(S/W)_t$ and the true S/W. This distribution shrinks as time passes. The difference between its 5th and 95th percentiles is 4% after 10 years and 2% after 50 years. After 10 years, the probability that the equilibrium S/W differs from the true S/W by at least 0.01 is 18% and the probability of at least a 0.05 difference is just under 3%. After 50 years, these probabilities are smaller, 9% and 1%, respectively. These results show that most of the time, investors gradually converge to the true optimal allocation, but the convergence can be slow. Slow convergence is common especially when b is low, as described next.

The left panels of Figure 6 show examples of learning paths when b is low. The first major difference from the case of high b is that it generally takes much longer for S/W to settle in a narrow range, if it settles at all during the first 300 years. For example, in the second panel on the left, S/W travels across the whole range of zero to one, and it continues moving even after 300 years. This difference is due to the fact that when b is close to 0, S/W has little effect on $\alpha = a - b(S/W)$. It is α, the conditional expected return, that investors learn about by observing realized returns. When $b \approx 0$, the variation in S/W does not cause much variation in realized

returns; the latter variation is mostly due to noise (x in equation (25)). Since realized returns do not help investors much in finding the optimal investment level, S/W keeps wandering around.

Another feature of low b is that S/W often settles at a level substantially different from the true S/W. For example, in the top left panel, S/W settles around 0.7, well above the true level of 0.5. To understand this result, recall that realized returns allow investors to learn about α that is conditional on the current level of S/W. If S/W were to stay constant forever, investors would eventually perfectly learn the value of α at that level of S/W. However, they would not learn a and b individually, so they would forever remain uncertain about α at any other level of S/W. This intuition helps us understand the path dependence in the left panels of Figure 6. After staying at a given level of S/W for a while, investors have learned more about α at that level of S/W than about α at any other level. As a result, they find it costly to change S/W because doing so would increase the uncertainty they face. Being stuck at a suboptimal level of S/W is costly as well, but the cost diminishes as b approaches zero. When b is close to zero, the cost of changing S/W may well exceed the cost of staying at a suboptimal level of S/W. In such cases, we observe S/W settling down at a level different from the true S/W, even after 300 years.

It would appear from Figure 6 that when b is low, investors can get stuck at the wrong investment level forever. They cannot, but convergence of S/W to 0.5 can take thousands of years.[14] To illustrate this fact, we run a single simulation exercise for one million years, using the true values of $a = 0.015$ and $b = 0.016$ (the 5th percentile of the prior distribution for b), which imply a true S/W of 0.9. We keep the same initial S/W of 0.9 as before, as well as the same priors. In this simulation, a few early negative return draws quickly push the equilibrium S/W down, and given the low value of b, S/W takes a long time to climb back up. We find that the equilibrium S/W is equal to 0.72 after 100 years, 0.77 after 500 years, and 0.78 after 1,000 years, well below the true value of 0.9. Even after 3,000 years, S/W is only 0.85. After 10,000 years, $S/W = 0.894$, and after a million years, S/W is only 0.0003 away from the true value. In short, convergence in S/W takes place eventually, but it can take so long that it is practically irrelevant. We conclude that when b is low, rational investors can get stuck at a suboptimal investment level. In other words, the equilibrium size of the active management industry can be suboptimal for a long period of time.

Let us briefly summarize the key findings from Figure 6. When b is high, investors find the optimal level of investment quickly. They learn a lot about a and b initially while S/W varies,

[14] To see that convergence to a different value cannot occur, note that at any interior value to which S/W converges, (29) holds. After infinitely many realizations of returns at a given S/W, there is no uncertainty about alpha at that value of S/W, so that $\sigma_\alpha = 0$. As a result, $\mathrm{E}(r_A|D) = \alpha$, the true value of alpha at that S/W, and $\mathrm{Var}(r_A|D) = \sigma_x^2$. Equation (29) then implies that S/W converges to $\alpha/(\gamma\sigma_x^2)$. It then follows from equations (31) and (32) that the value to which S/W converges must be the true value, $a/(b + \gamma\sigma_x^2)$.

but their learning all but stops after S/W settles down at or near the true S/W. When b is low, learning is highly path-dependent. S/W fluctuates much longer before it settles in a narrow range, if it settles at all. This narrow range need not include the true S/W, and investors can get stuck at a suboptimal investment level for a very long time.

In our final analysis, we let our investors learn from the year-by-year series of actual mutual fund returns. We use Prior 2 ($b \geq 0$) and the same data as in Figure 4. The results are plotted in Figure 7. The first three panels plot the time series of the posterior distributions of a, b, and α. All of these distributions shrink as time passes, but the distributions of a and b shrink only modestly compared to the distribution of α. By 2006, investors are confident that the annual α at the current S/W is within 1% of zero, but they remain quite uncertain about a and b. As before, we see that learning about decreasing returns to scale is slow. Some insight into the low speed of learning is provided by the learning path plotted in the bottom right panel of Figure 7. That plot is similar to the examples in Figure 6 (middle column, lower rows) in which b is at its median prior value and the true $S/W = 0.7$. Many of the S/W values cluster in the neighborhood of 0.7. As a result, when a regression line is fitted through the scatter of points, the intercept and slope clearly remain very uncertain even after observing 44 years of data.

5. Relation to Berk and Green (2004)

A central feature of our model is that active managers face decreasing returns to scale in their abilities to generate alpha. In this respect our approach follows Berk and Green (2004), but there are important differences. First, Berk and Green (hereafter BG) assume that decreasing returns apply at the level of individual funds, whereas we assume they apply to the active management industry as a whole. That is, we assume an individual fund's alpha is decreasing in the total amount invested by all active funds.[15] It seems reasonable that even a small fund finds it more difficult to identify profitable investment opportunities as the overall amount of actively-invested capital grows and thereby moves prices to eliminate such opportunities.[16] Assuming decreasing returns at the individual fund level seems plausible as well, though it encounters the question of

[15] It is easy to show that our assumption of decreasing returns to scale at the aggregate level also implies decreasing returns to scale at the individual fund level. However, this implication weakens as the number of funds grows larger. Empirical evidence indicating decreasing returns to scale at the fund level, especially among small-cap mutual funds, is provided by Chen, Hong, Huang, and Kubik (2004) and Pollet and Wilson (2008). Related evidence for hedge funds, at the fund level as well as aggregate level, is provided by Fung, Hsieh, Naik, and Ramadorai (2008).

[16] A similar perspective is adopted by Glode and Green (2010) who argue that fund returns can be decreasing in the size of a sector or trading strategy, as well as in the size of the fund itself. Glode and Green develop a model of information spillovers that can rationalize performance persistence in hedge funds.

what happens if multiple funds merge or additional managers are hired. Presumably, in the absence of aggregate effects, such mergers or hires would simply keep increasing the fund size at which decreasing returns take their bite.

A second difference in our treatment of decreasing returns to scale is that we do not assume that investors know the degree to which alpha drops as the amount of active management increases. In our parameterization of decreasing returns in (5), the values of both a and b are unknown. In contrast, the model in BG corresponds to a setting in which a is unknown but b is known.[17] As discussed earlier, when both a and b in (5) are unknown, investors face an interesting learning problem in which the true values of those parameters are never fully learned.

Another difference from BG is that their investors face $\tilde{\alpha} = 0$, whereas our investors perceive $\tilde{\alpha} > 0$. We solve for the Nash equilibrium among investors maximizing (10). BG do not solve the investors' optimization problem explicitly; instead, they fix $\tilde{\alpha} = 0$ by invoking the assumption that non-benchmark risk can be completely diversified away across many funds. BG argue that if a large number of funds were to have positive alphas, one could combine them in a portfolio with a positive alpha and zero non-benchmark risk; $\tilde{\alpha} = 0$ is therefore a necessary condition for equilibrium. Recall from Proposition 1 that our model implies $\tilde{\alpha} = 0$ as well if investors are risk neutral and the number of investors is infinite. With a finite number of investors, however, $\tilde{\alpha} > 0$ because investors internalize some of the reduction in alpha caused by their own investment.

Even with an infinite number of investors, $\tilde{\alpha} > 0$ if investors are risk-averse because they then require compensation for both non-diversifiable risk ($\sigma_x > 0$) and uncertainty about α ($\sigma_\alpha > 0$). These effects are clear from equation (32), which applies when $N \to \infty$ and $M \to \infty$. However, we do not wish to leave readers with the impression that alpha in that setting is necessarily large. Equations (31) and (32) imply that $\tilde{\alpha} = (S/W)\gamma(\sigma_x^2 + \sigma_\alpha^2)$. This value is small once learning proceeds to the point where $\sigma_\alpha \approx 0$. Even with $S/W = 1$, the values of γ and σ_x specified in our numerical investigation (2 and 0.02, respectively) then imply a value of $\tilde{\alpha}$ equal to only 8 basis points per annum. Uncertainty about α increases the equilibrium value of $\tilde{\alpha}$, but only slightly unless learning is just beginning. Thus, even though our modeling of the determinants of equilibrium alpha is rather different from that of BG, their zero-alpha condition is not at sharp odds, in practical terms, with a setting in which $\sigma_x > 0$, $\sigma_\alpha > 0$, and there are many funds and investors.[18]

[17] BG denote the quantity corresponding to our "b" as "a" in their quadratic parameterization, and they view this quantity as known. Their "α" corresponds to our "a"—they use "α" to denote the expected return gross of fees and costs, whereas we use "α" to denote the expected benchmark-adjusted return received by investors (see equation (2)).

[18] A closely related statement is that in our model, past performance predicts future performance, but only slightly.

In their diversification argument justifying zero alpha, BG rely on the presence of many funds. This assumption is at some tension with BG's treatment of fund managers as monopolists. In the BG model, each manager sets a proportional fee rate by taking into account its effect on the amount of assets under management. That amount ends up maximizing expected profit received in total by managers and investors; the analogous aggregate amount in our setting is given in equation (18).[19] In our model, with multiple competing funds, that discretionary component of the fee disappears ($f = 0$), and managers become price takers with respect to their equilibrium fees. When there are many competing investors as well, the amount invested under risk neutrality is twice as large as in the BG model; investment reaches the level that produces zero expected profit.

The specification that brings our model closest to that of BG involves a single manager ($M = 1$) and many risk-neutral investors ($N \to \infty$, $\gamma = 0$). With $M = 1$, we obtain $f = \tilde{a}/2$, as in BG.[20] With $N \to \infty$, the externality present with fewer investors, which is absent from BG, disappears. With $\gamma = 0$, we obtain BG's zero-alpha condition because there is no compensation for risk. Equation (18) shows that equilibrium under the above specification produces a profit-maximizing size of the industry that is analogous to the profit-maximizing fund size obtained in BG.

6. Conclusion

It seems puzzling that active management remains popular despite its track record. We propose a potential resolution to this puzzle. In a model with competing investors and fund managers, we find that the equilibrium size of the active management industry can be large even after a significantly negative track record. The key to this result is the belief that active managers face decreasing returns to scale. If investors instead believed that returns to scale were constant, they would allocate nothing to active management even if they were initially more optimistic about active managers' abilities.

Under decreasing returns to scale, investors adjust their allocation in response to performance until the expected return going forward is sufficiently attractive. Given the observed year-by-year performance of active mutual funds over the past four decades, our model makes two predictions about the investors' proportional allocation to active management: this allocation should have decreased over time, but it should also remain substantial. Consistent with the first prediction,

[19] Our equation (18) corresponds to BG's equation (26), $q_t^*(\phi_t) = \phi_t/2a$. BG's a corresponds to our $\tilde{b}$, their ϕ_t corresponds to our $\tilde{a}$, and their q_t^* corresponds to our expected-profit-maximizing S/W.

[20] Here we refer to the special case of BG in which the profit/cost function is quadratic, as it is in our model. BG analyze not only this special case but also the more general case of convex costs.

passive investing has grown dramatically since its humble beginnings in the 1970s. Consistent with the second prediction, active investing remains more popular than passive investing.

Investors in our model face endogeneity that limits their learning—what they learn affects how much they allocate to active management, but what they allocate affects how much they learn. The equilibrium allocation typically ceases to fluctuate after just a few years, at which point learning about returns to scale essentially stops. As a result, investors never accurately learn the degree of decreasing returns to scale. We also find that when active returns are not very sensitive to the industry's size, this size can fluctuate at suboptimal levels for a long time.

Future research can explore additional aspects of learning about parameters governing returns to scale. These parameters are held constant in our model, for simplicity, but they could plausibly vary due to exogenous shocks. For example, shocks to liquidity would likely induce changes in the degree of decreasing returns to scale. In such a setting, parameter uncertainty gets refreshed every so often, so that learning is always at a relatively early stage. The probability that the industry size is suboptimal at any point in time is then higher than in the constant-parameter framework, and so is the probability of observing unusually large positive or negative t-statistics. Future work could also further explore the economic importance of the incomplete learning about returns to scale. We have a lot yet to learn about learning in active management.

Appendix

In this appendix, we derive Propositions 1 and 2. After first justifying the objective in (10), we analyze the risk-neutral setting in which $\gamma = 0$ and then turn to the risk-averse setting with $\gamma > 0$.

A.1. Objective function

The total return on investor j's portfolio is given by

$$\begin{aligned} R_j &= (1 - \delta_j'\iota_N - \phi_j'\iota_K)R_f + \delta_j' R_F + \phi_j' R_B \\ &= R_f + \delta_j'(R_F - \iota_N R_f) + \phi_j'(R_B - \iota_K R_f) \\ &= R_f + \delta_j' r_F + \phi_j' r_B \end{aligned} \tag{A1}$$

where R_f denotes the interest rate, R_F and R_B denote total rates of return on the N funds and K benchmarks, r_F and r_B denote returns in excess of the interest rate, ι_n denotes an n-vector of 1's, and δ_j and ϕ_j denote the vectors of weights on the funds and benchmarks. (Note that the weights across all assets sum to 1 by construction.) We assume investor j solves the problem

$$\max_{\delta_j, \phi_j} \left\{ E(R_j|D) - \frac{\gamma}{2}\text{Var}(R_j|D) \right\}, \tag{A2}$$

where the allocations to the benchmarks and riskless asset are unrestricted. Combining (A1) with equations (2) and (3) gives

$$\begin{aligned} R_j &= R_f + \delta_j'(\underline{\alpha} + Br_B + u) + \phi_j' r_B \\ &= R_f + \delta_j' r + (\delta_j' B + \phi_j')r_B \\ &= R_f + \delta_j' r + \psi_j' r_B, \end{aligned} \tag{A3}$$

where

$$\psi_j = B'\delta_j + \phi_j. \tag{A4}$$

From (A3) we obtain

$$\text{E}(R_j) = R_f + \delta_j'\text{E}(r|D) + \psi_j'\text{E}(r_B|D) \tag{A5}$$

$$\text{Var}(R_j) = \delta_j'\text{Var}(r|D)\delta_j + \psi_j'\text{Var}(r_B|D)\psi_j, \tag{A6}$$

recalling that r and r_B are uncorrelated. Because we do not constrain ϕ_j, the weights on the benchmarks, we see from (A4) through (A6) that the maximization in (A2) is equivalent to

$$\max_{\delta_j, \psi_j} \left\{ R_f + \delta_j'\text{E}(r|D) + \psi_j'\text{E}(r_B|D) - \frac{\gamma}{2}\left(\delta_j'\text{Var}(r|D)\delta_j + \psi_j'\text{Var}(r_B|D)\psi_j\right) \right\}. \tag{A7}$$

We then see that (A7) separates into two maximation problems: one is given by

$$\max_{\psi_j} \left\{ \psi_j' \mathrm{E}(r_B|D) - \frac{\gamma}{2} \psi_j' \mathrm{Var}(r_B|D) \psi_j \right\}, \tag{A8}$$

and the other—the relevant problem for our purposes—is given by (10).

A.2. Risk Neutrality

Let s denote the $M \times 1$ vector whose i-th element is s_i. Observe that

$$s = \frac{W}{N} \sum_{j=1}^{N} \delta_j \tag{A9}$$

$$S = \iota_M' s = \frac{W}{N} \sum_{j=1}^{N} \iota_M' \delta_j, \tag{A10}$$

where ι_M is an $M \times 1$ vector of ones. The vector of benchmark-adjusted fund returns is given by

$$r = a\iota_M - b\frac{S}{W}\iota_M - \underline{f} + u, \tag{A11}$$

so expected fund returns can be written as

$$\mathrm{E}(r|D) = \tilde{a}\iota_M - \frac{\tilde{b}}{N} \sum_{j=1}^{N} \iota_M' \delta_j \iota_M - \underline{f} \tag{A12}$$

where $\underline{f}$ is an $M \times 1$ vector whose i-th element is f_i. Substituting into (10), setting $\gamma = 0$, gives investor j's problem as

$$\max_{\delta_j} \left\{ \delta_j' \left(\tilde{a}\iota_M - \frac{\tilde{b}}{N} \sum_{n \neq j} \iota_M' \delta_n \iota_M - \underline{f} \right) - \delta_j' \frac{\tilde{b}}{N} \iota_M \iota_M' \delta_j \right\}, \tag{A13}$$

subject to the restrictions

$$\iota_M' \delta_j \leq \delta^* \tag{A14}$$

$$\delta_{i,j} \geq 0 \ \ \forall i, j, \tag{A15}$$

where $\delta_{i,j}$ denotes the i-th element of δ_j. We here impose the leverage constraint in (A14) from the outset and then consider the separate cases where it does and does not bind.

In a Nash equilibrium, wherein each investor takes the optimal decisions of other investors as given, investor j's first-order condition from (A13) is given by

$$\tilde{a}\iota_M - \underline{f} - \frac{\tilde{b}}{N} \iota_M \iota_M' (\sum_{n=1}^{N} \delta_n + \delta_j) - \lambda_1 \iota_M - \lambda_2 = 0, \tag{A16}$$

where the scalar λ_1 and the $M \times 1$ vector λ_2 contain the multipliers associated with the constraints in (A14) and (A15). All N investors are identical, and we confine attention to symmetric equilibria,

$$\delta_j = \underline{\delta}, \qquad j = 1, \ldots, N. \tag{A17}$$

Imposing (A17) on the first-order conditions in (A16) then implies

$$\begin{aligned} 0 &= \tilde{a}\iota_M - \underline{f} - \frac{(N+1)}{N}\tilde{b}\iota_M\iota_M'\delta - \lambda_1\iota_M - \lambda_2 \\ &= \left(\tilde{a} - \frac{(N+1)}{N}\tilde{b}\,(\iota_M'\delta) - \lambda_1\right)\iota_M - \underline{f} - \lambda_2. \end{aligned} \tag{A18}$$

From (A18) we see that for any δ satisfying (A18), all other values of δ giving the same value of $\iota_M'\delta$ also satisfy (A18). We thus define the scalar,

$$\bar{\delta} = \frac{1}{M}\iota_M'\delta. \tag{A19}$$

We also see from (A18) that, for all funds receiving a positive investment, the elements of $\underline{f}$ are equal to a common scalar value f, since the corresponding elements of λ_2 for those funds are equal to zero. If all M funds receive some positive investment, then

$$\underline{f} = f\iota_M. \tag{A20}$$

From (A18) through (A20) we thus obtain the condition that, when all funds receive positive investment and the leverage constraint in (A14) does not bind,

$$\bar{\delta} = \frac{\tilde{a} - f}{\tilde{b}}\left(\frac{1}{M}\right)\left(\frac{N}{N+1}\right), \tag{A21}$$

or

$$\frac{S}{W} = M\bar{\delta} = \frac{\tilde{a} - f}{\tilde{b}}\left(\frac{N}{N+1}\right). \tag{A22}$$

When $\tilde{a} \leq 0$, then $S/W = 0$.

The M managers, aware of the above equilibrium conditions, set their fees before investors make their decisions. If $M = 1$, then the monopolistic manager chooses f so that his resulting equilibrium fund size s maximizes fee revenue,

$$fs = f\frac{W}{N}\sum_{j=1}^{N}\delta_j = f\frac{W}{N}N\delta = fW\bar{\delta} = fW\frac{\tilde{a} - f}{\tilde{b}}\left(\frac{N}{N+1}\right), \tag{A23}$$

which gives his optimal fee when $\tilde{a} > 0$ as

$$f = \frac{\tilde{a}}{2}. \tag{A24}$$

If $\tilde{a} < 0$, then the fund receives no investment for any non-negative fee. When $M > 1$, competition among the (non-cooperative) multiple managers results in

$$f = 0. \tag{A25}$$

To see this, suppose instead that two or more funds receive positive investment with $f > 0$. If any manager then lowers his fee infinitesimally below f, the risk neutral investors would simply transfer all investments currently allocated to the other funds to that lower-fee fund.

Combining the above observations about f with (A22) gives the equilibrium allocation to active management. When $\tilde{a} > 0$ the equilibrium allocation to active management is

$$\frac{S}{W} = \left(\frac{1}{2}\right)\frac{\tilde{a}}{\tilde{b}}\left(\frac{N}{N+1}\right) \quad \text{for } M = 1, \tag{A26}$$

and

$$\frac{S}{W} = \frac{\tilde{a}}{\tilde{b}}\left(\frac{N}{N+1}\right) \quad \text{for } M > 1. \tag{A27}$$

When the right-hand side of either (A26) or (A27) exceeds δ^*, so that the leverage constraint in (A14) binds, then $S/W = \delta^*$. In the latter case, when $M = 1$, the manager relies on (A21) to set

$$f = \tilde{a} - \delta^* \tilde{b}\left(\frac{N+1}{N}\right), \tag{A28}$$

which then exceeds the value in (A24). The value of f is still zero for $M > 1$, for the same reason given earlier.

The above analysis also implies that, when $\tilde{a} > 0$, $E(r|D) = \tilde{\alpha}\iota_M$. When the leverage constraint in (A14) does not bind,

$$\tilde{\alpha} = \left(\frac{1}{2}\right)\frac{\tilde{a}}{N+1} \quad \text{for } M = 1, \tag{A29}$$

and

$$\tilde{\alpha} = \frac{\tilde{a}}{N+1} \quad \text{for } M > 1. \tag{A30}$$

When the leverage constraint binds,

$$\tilde{\alpha} = \left(\frac{1}{N}\right)\tilde{b}\delta^* \quad \text{for } M = 1, \tag{A31}$$

and

$$\tilde{\alpha} = \tilde{a} - \tilde{b}\delta^* \quad \text{for } M > 1. \tag{A32}$$

When the leverage constraint binds and there are competing managers, then investors earn a non-trivial positive $\tilde{\alpha}$ even for large N. Otherwise, $\alpha \to 0$ as $N \to \infty$.

A.3. Risk Aversion

The benchmark-adjusted fund returns in (A11) can be written as

$$\begin{aligned} r &= \tilde{a}\iota_M - \tilde{b}\frac{S}{W}\iota_M - \underline{f} + u + (a-\tilde{a})\iota_M - (b-\tilde{b})\frac{S}{W}\iota_M \\ &= \tilde{a}\iota_M - \tilde{b}\frac{1}{N}\sum_{n=1}^{N}\iota_M'\delta_n\iota_M - \underline{f} + \left\{u + (a-\tilde{a})\iota_M - (b-\tilde{b})\frac{1}{N}\sum_{n=1}^{N}\iota_M'\delta_n\iota_M\right\}, \end{aligned}$$

and taking the variance gives

$$\begin{aligned} \text{Var}(r|D) &= \underbrace{\sigma_x^2\iota_M\iota_M' + \sigma_\epsilon^2 I_M}_{\sigma_u^2} + \sigma_a^2\iota_M\iota_M' - 2\sigma_{ab}(\frac{1}{N}\sum_{n=1}^{N}\iota_M'\delta_n)\iota_M\iota_M' + \sigma_b^2\frac{1}{N^2}\left[\sum_{n=1}^{N}\iota_M'\delta_n\right]^2\iota_M\iota_M' \\ &= \sigma_1^2\iota_M\iota_M' + \sigma_\epsilon^2 I_M - 2\sigma_{ab}(\frac{1}{N}\sum_{n=1}^{N}\iota_M'\delta_n)\iota_M\iota_M' + \sigma_b^2\frac{1}{N^2}\left[\sum_{n=1}^{N}\iota_M'\delta_n\right]^2\iota_M\iota_M', \end{aligned}$$

where

$$\sigma_1^2 = \sigma_x^2 + \sigma_a^2. \tag{A33}$$

When facing the problem in (10), each investor j recognizes that, since all funds are identical, the solution will be of the form $\delta_j = \delta_{(j)}\iota_M$, where $\delta_{(j)}$ is a scalar. Investors also recognize that since they are all identical, the other $N-1$ investors will all have solutions of the form $\delta_n = \delta^*\iota_M$, where δ^* is a scalar. As a result, we can write

$$\sum_{n=1}^{N}\iota_M'\delta_n = \iota_M'\left[\delta_{(j)}\iota_M + (N-1)\delta^*\iota_M\right] = M\left(\delta_{(j)} + (N-1)\delta^*\right).$$

Since it is known to managers and investors that the values of a and b are identical across funds, we assume that investors face fees of the form $\underline{f} = f\iota_M$, where f is a scalar. Therefore, each investor j solves for $\delta_{(j)}$ that maximizes the quantity

$$\begin{aligned} &\delta_j'\text{E}(r|D) - \frac{\gamma}{2}\delta_j'\text{Var}(r|D)\delta_j \\ &= \delta_{(j)}\iota_M'\text{E}(r|D) - \frac{\gamma}{2}\delta_{(j)}^2\iota_M'\text{Var}(r|D)\iota_M \\ &= \delta_{(j)}\iota_M'\left[\iota_M(\tilde{a}-f) - \tilde{b}\frac{1}{N}\iota_M M\left(\delta_{(j)} + (N-1)\delta^*\right)\right] - \\ &\quad \frac{\gamma}{2}\delta_{(j)}^2\iota_M'\left[\sigma_1^2\iota_M\iota_M' + \sigma_\epsilon^2 I_M - 2\sigma_{ab}\frac{M}{N}\left(\delta_{(j)} + (N-1)\delta^*\right)\iota_M\iota_M' + \sigma_b^2\frac{M^2}{N^2}\left(\delta_{(j)} + (N-1)\delta^*\right)^2\iota_M\iota_M'\right]\iota_M \end{aligned}$$

subject to the constraints (A14) and (A15). This is equivalent to maximizing

$$\begin{aligned} &\delta_{(j)}M(\tilde{a}-f) - \delta_{(j)}^2\tilde{b}\frac{M^2}{N} - \delta_{(j)}\tilde{b}\frac{M^2(N-1)}{N}\delta^* \\ &-\frac{\gamma}{2}\delta_{(j)}^2M^2\sigma_1^2 - \frac{\gamma}{2}\delta_{(j)}^2M\sigma_\epsilon^2 + \gamma\delta_{(j)}^3\sigma_{ab}\frac{M^3}{N} + \gamma\delta_{(j)}^2\sigma_{ab}\frac{M^3(N-1)}{N}\delta^* \\ &-\frac{\gamma}{2}\delta_{(j)}^4\sigma_b^2\frac{M^4}{N^2} - \gamma\delta_{(j)}^3\sigma_b^2\frac{M^4(N-1)}{N^2}\delta^* - \frac{\gamma}{2}\delta_{(j)}^2\sigma_b^2\frac{M^4(N-1)^2}{N^2}(\delta^*)^2 - \lambda_1\left(M\delta_{(j)} - \bar{\delta}\right) - \lambda_2\delta_{(j)}. \end{aligned}$$

Taking the first derivative with respect to $\delta_{(j)}$, we obtain the first-order condition:

$$\begin{aligned}
0 &= M(\tilde{a}-f) - 2\delta_{(j)}\tilde{b}\frac{M^2}{N} - \tilde{b}\frac{M^2(N-1)}{N}\delta^* \\
&\quad -\gamma\delta_{(j)}M^2\sigma_1^2 - \gamma\delta_{(j)}M\sigma_\epsilon^2 + 3\gamma\delta_{(j)}^2\sigma_{ab}\frac{M^3}{N} + 2\gamma\delta_{(j)}\sigma_{ab}\frac{M^3(N-1)}{N}\delta^* \\
&\quad -2\gamma\delta_{(j)}^3\sigma_b^2\frac{M^4}{N^2} - 3\gamma\delta_{(j)}^2\sigma_b^2\frac{M^4(N-1)}{N^2}\delta^* - \gamma\delta_{(j)}\sigma_b^2\frac{M^4(N-1)^2}{N^2}(\delta^*)^2 - M\lambda_1 - \lambda_2.
\end{aligned}$$

Dividing through by M and recognizing that, in equilibrium, $\delta_{(j)} = \delta^*$ for all j, we have

$$\begin{aligned}
0 &= \tilde{a} - f - \lambda_1 - \frac{\lambda_2}{M} - 2\delta^*\tilde{b}\frac{M}{N} - \tilde{b}\frac{M(N-1)}{N}\delta^* \\
&\quad -\gamma\delta^*M\sigma_1^2 - \gamma\delta^*\sigma_\epsilon^2 + 3\gamma\delta^{*2}\sigma_{ab}\frac{M^2}{N} + 2\gamma\delta^{*2}\sigma_{ab}\frac{M^2(N-1)}{N} \\
&\quad -2\gamma\delta^{*3}\sigma_b^2\frac{M^3}{N^2} - 3\gamma\delta^{*3}\sigma_b^2\frac{M^3(N-1)}{N^2} - \gamma\delta^{*3}\sigma_b^2\frac{M^3(N-1)^2}{N^2} \\
&= \tilde{a} - f - \lambda_1 - \frac{\lambda_2}{M} - M\delta^*\left[\frac{\tilde{b}(N+1)}{N} + \gamma\sigma_1^2 + \frac{\gamma\sigma_\epsilon^2}{M}\right] + (M\delta^*)^2\frac{\gamma\sigma_{ab}}{N}[2N+1] \\
&\quad -(M\delta^*)^3\frac{\gamma\sigma_b^2(N+1)}{N}.
\end{aligned}$$

Note that

$$\frac{S}{W} = \iota_M'\frac{s}{W} = \iota_M'\frac{1}{N}\sum_{n=1}^{N}\delta_n = \iota_M'\frac{1}{N}N\delta^*\iota_M = M\delta^*. \tag{A34}$$

As $M \to \infty$ and $N \to \infty$, the first-order condition then becomes, using (A34),

$$0 = \tilde{a} - f - \lambda_1 - \frac{S}{W}\left[\tilde{b} + \gamma\sigma_1^2\right] + \left(\frac{S}{W}\right)^2 2\gamma\sigma_{ab} - \left(\frac{S}{W}\right)^3\gamma\sigma_b^2. \tag{A35}$$

Following the earlier discussion, we set $f = 0$ when the number of funds (M) is infinite. When the constraint in (A15) does not bind and thus $\lambda_1 = 0$, (A35) is identical to equation (21) in Proposition 2, noting (A33). It can be verified that this equation has one positive real solution for S/W. If that solution exceeds δ^*, then $S/W = \delta^*$.

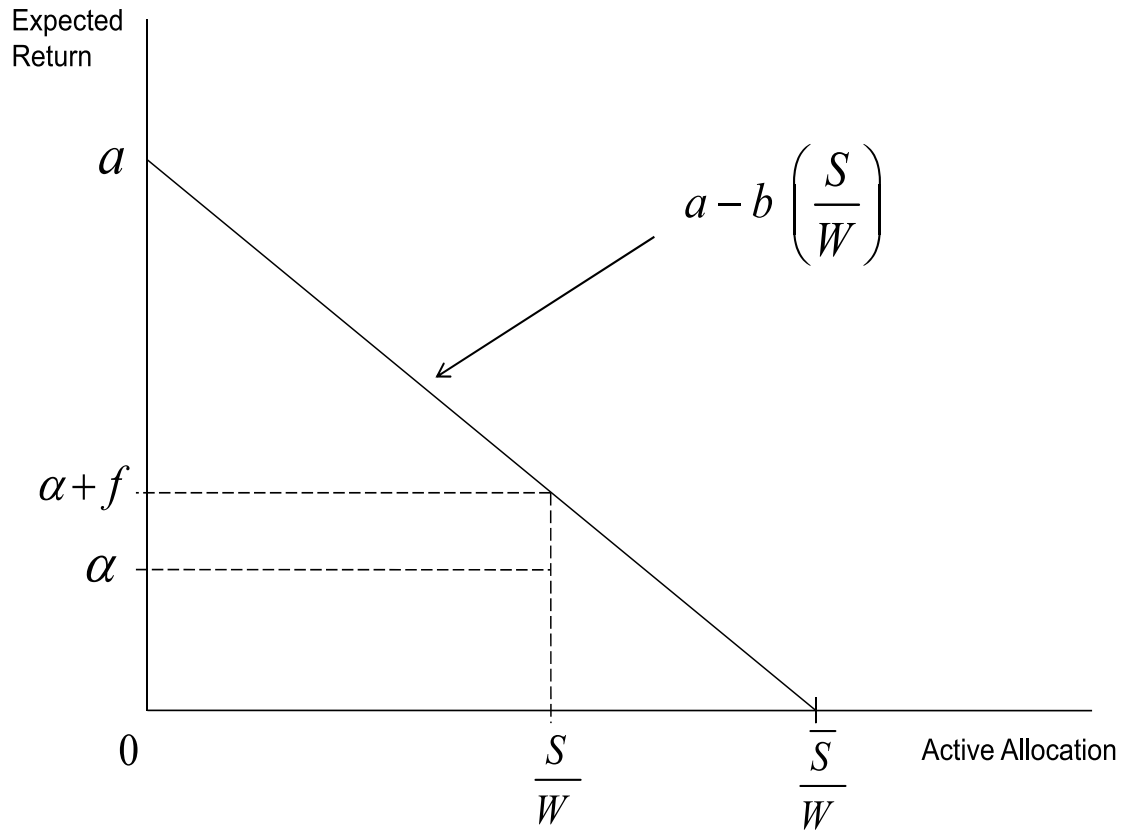

Figure 1. Decreasing returns to scale in the active management industry. This figure plots the theoretical relation between the expected benchmark-adjusted excess fund return before fees against the relative size of the active management industry. Specifically, it plots equation (9): $\alpha + f = a - b\frac{S}{W}$, where α is the expected benchmark-adjusted excess fund return earned by investors, f is the proportional fee charged by the fund manager, S is the aggregate size of the active management industry, and W is the investors' total investable wealth. As long as $b > 0$, the industry exhibits decresing returns to scale. The values of α, f, and S are determined in equilibrium. At $S = \overline{S}$, we have $\alpha = f = 0$.

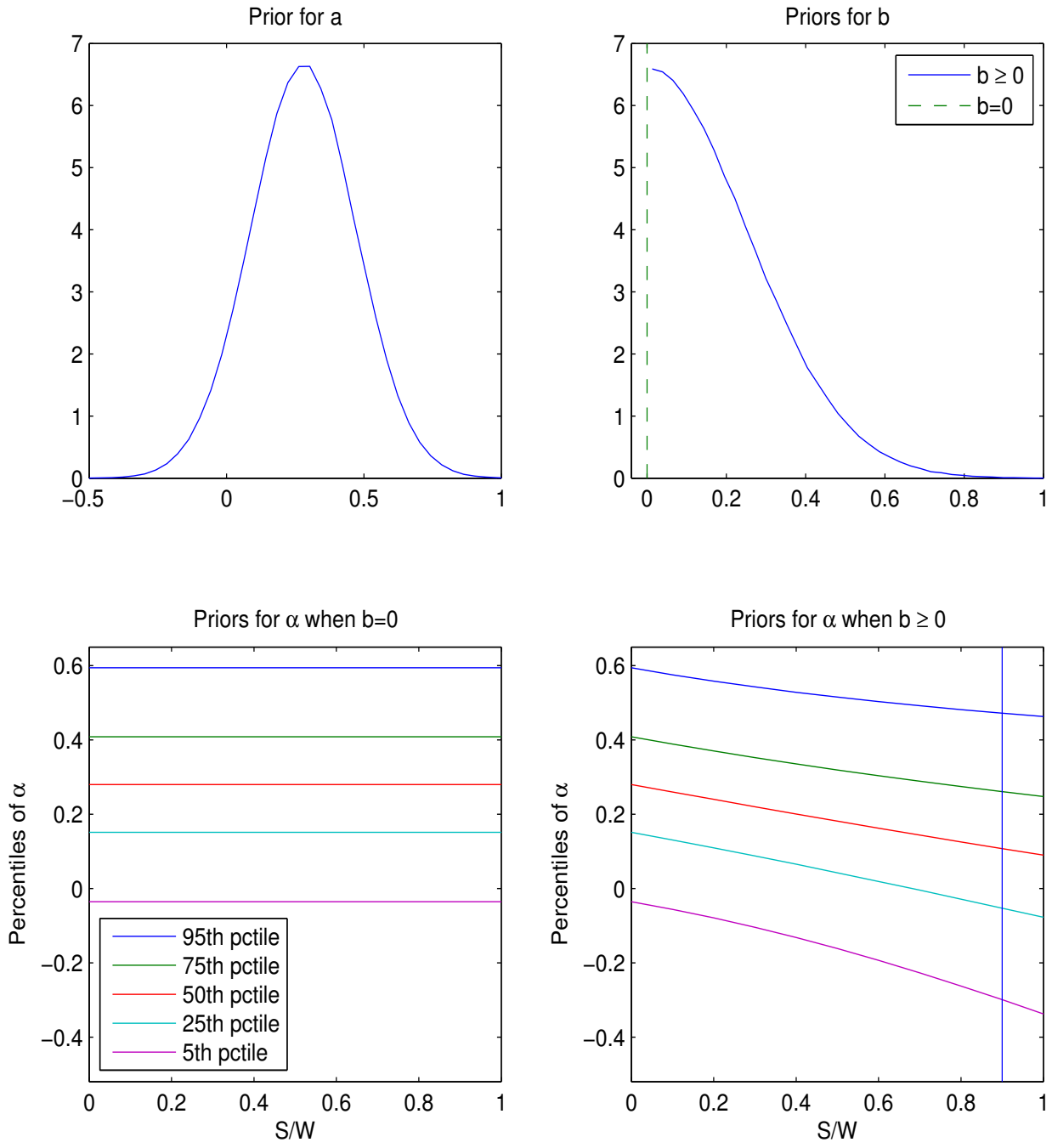

Figure 2. Prior distributions. This figure plots the prior distributions for the parameters of the function in equation (9). Panel A plots the prior for a, which is normal with the mean of 0.28 and standard deviation of 0.19. Panel B plots two different prior distributions for b: $b = 0$ (constant returns to scale), and $b \geq 0$ (decreasing returns to scale). The former prior is a spike at $b = 0$. The latter prior is truncated normal with the mode of zero, mean of 0.2, and standard deviation of 0.15. Under this prior, the initial equilibrium allocation to active management is $S/W = 0.9$. The parameters a and b are independent a priori. Panels C and D plot the 5th, 25th, 50th, 75th, and 95th percentiles of the implied prior distributions for $\alpha = a - b(S/W)$ as a function of S/W (in the competitive case with $f = 0$). Panel C corresponds to the prior $b = 0$, for which the distribution of α is invariant to S/W. Panel D corresponds to the prior $b \geq 0$, for which the distribution of α shifts toward smaller values as S/W increases.

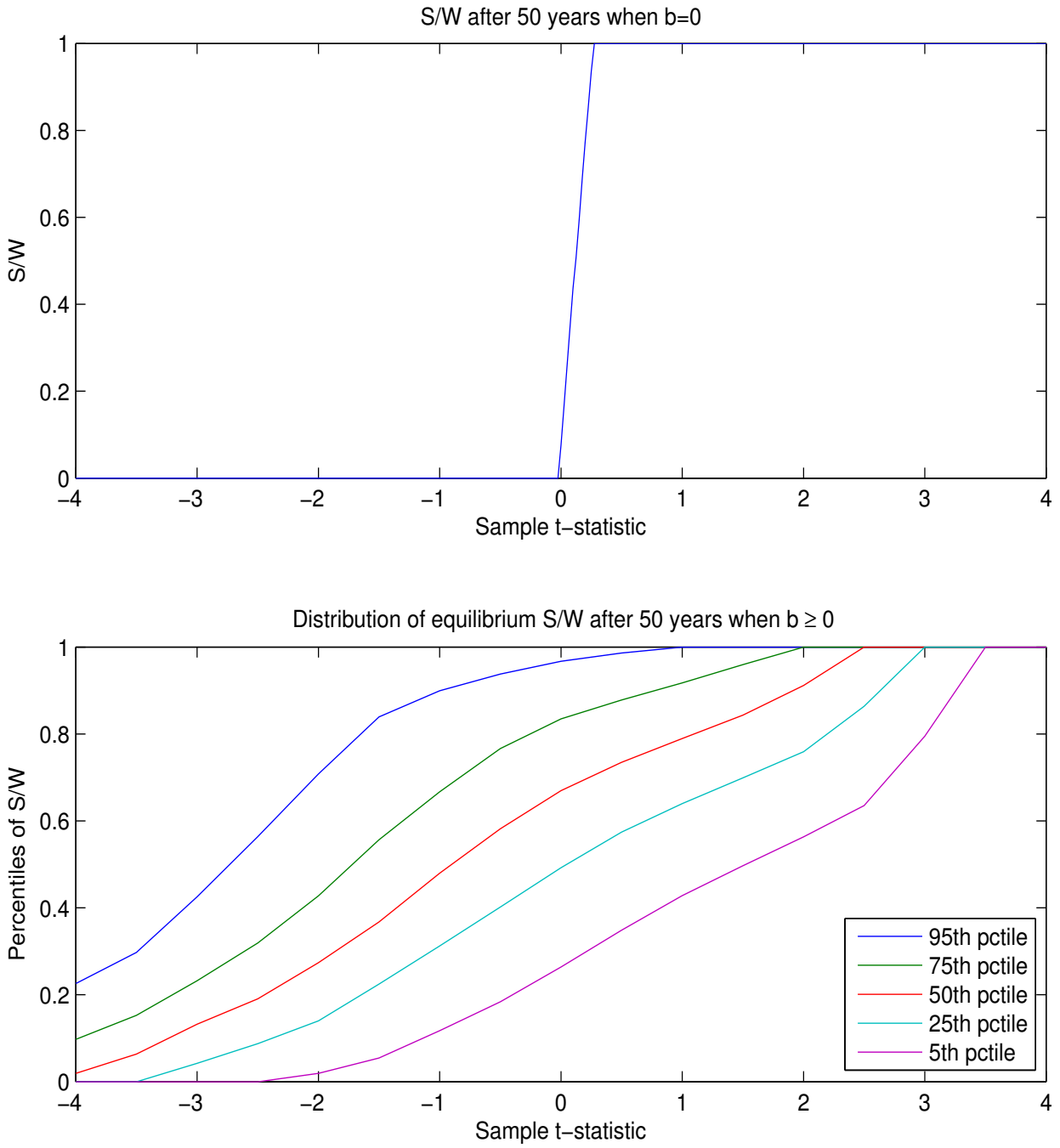

Figure 3. The posterior distribution of the equilibrium allocation to active management conditional on various levels of overall past performance. This figure plots selected percentiles of the posterior distribution of S/W, the equilibrium allocation to active management, conditional on the t-statistic associated with the industry's historical alpha computed over a period of $T = 50$ years. Panel A corresponds to the prior $b = 0$ (constant returns to scale); the distribution of S/W then collapses into a single value because the t-statistic is a sufficient statistic for S/W. Panel B corresponds to the prior $b \geq 0$ (decreasing returns to scale). Note that when $b = 0$, investors observing negative past performance optimally choose to invest nothing in active management, but when $b \geq 0$, they invest substantial amounts.

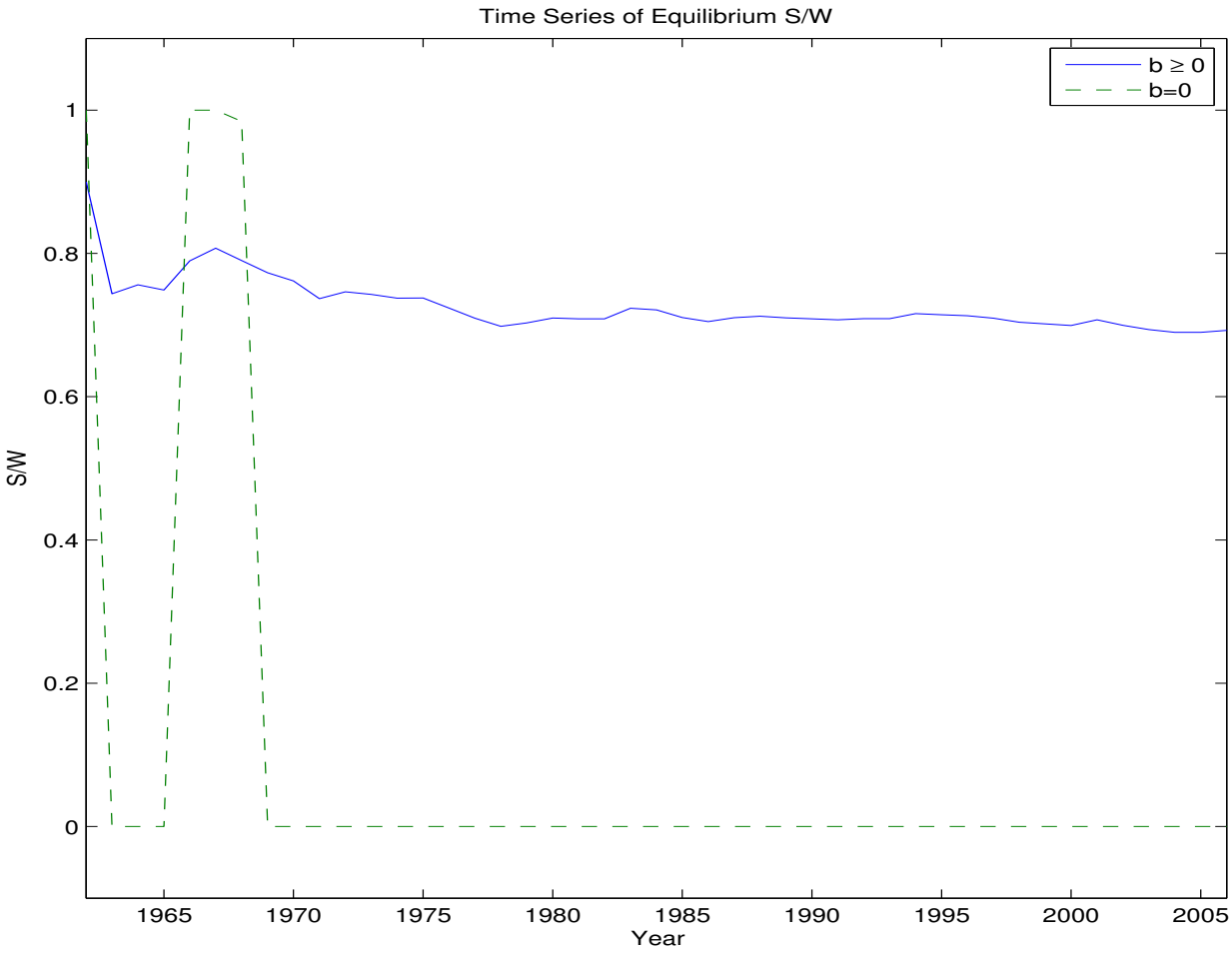

Figure 4. Equilibrium allocations to active management based on the year-by-year series of actual mutual fund returns. For each year from 1962 through 2006, the figure plots the equilibrium allocation to active management, S/W, computed given the previous histories of equilibrium allocations and actual returns on the aggregate portfolio of U.S. actively managed mutual funds. The fund returns are adjusted for exposures to the three Fama-French factors. The first year's allocation is based on the prior distribution. Two priors are considered: $b \geq 0$ (decreasing returns to scale) and $b = 0$ (constant returns to scale).

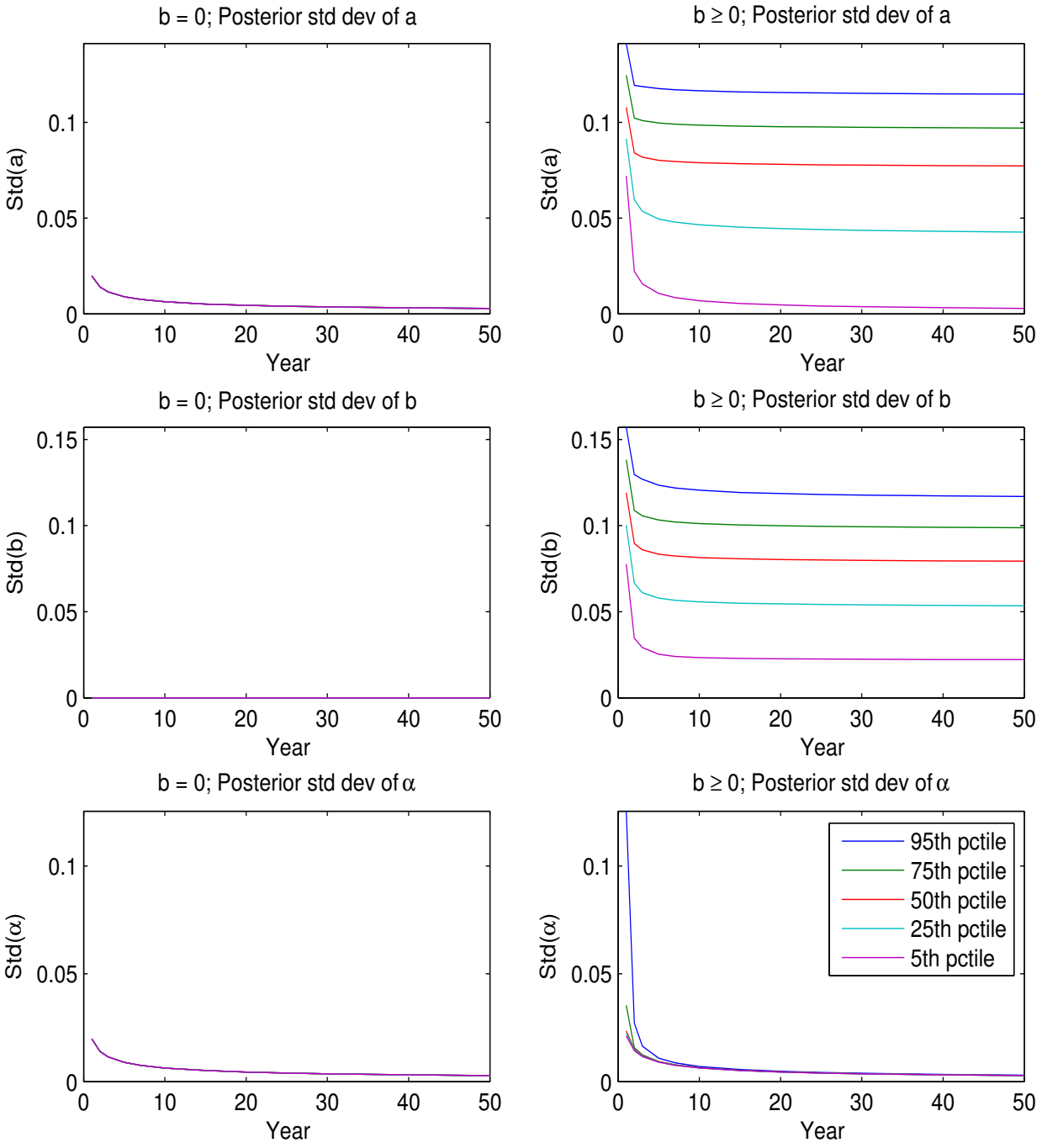

Figure 5. Posterior standard deviations. This figure plots the posterior standard deviations of a, b, and α as a function of time. The three panels on the left correspond to the prior $b = 0$ (constant returns to scale); the three panels on the right represent the prior $b \geq 0$ (decreasing returns to scale). Each panel on the right plots selected percentiles of the distribution of the given standard deviation across many simulated samples. Under the prior $b = 0$, there is no dispersion in this distribution, so the three panels on the left plot single lines. Also when $b = 0$, a and α coincide, so the top and bottom left panels look identical. The middle left panel looks empty because the posterior standard deviation of b is zero when $b = 0$.

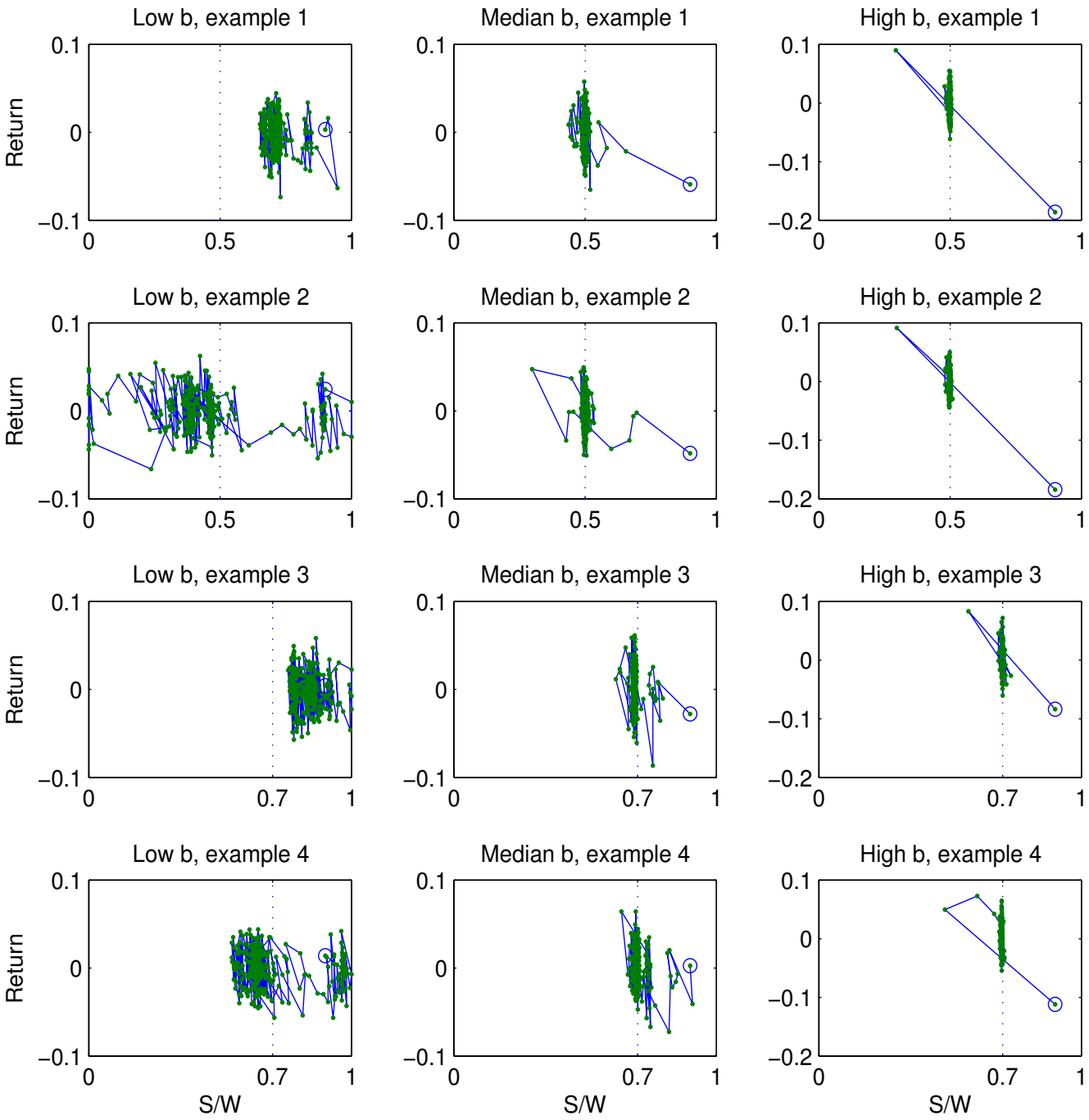

Figure 6. Examples of learning paths under decreasing returns to scale. This figure plots representative examples of learning paths for various random samples under the $b \geq 0$ prior. Each of the 12 panels plots aggregate active fund returns $r_{A,t}$ against the aggregate allocation to the active industry $(S/W)_t$ for $t = 1, \ldots, 300$ years. The three columns of panels correspond to three different values of b: "low" (5th percentile of the prior distribution, 0.02), "median" (50th percentile, 0.17), and "high" (95th percentile, 0.49). Given the value of b, the value of a is computed so that the true S/W is either $S/W = 0.5$ (first two rows of panels) or $S/W = 0.7$ (last two rows). The (a, b) pair is then used to generate random samples of active returns, which are then used to update the $b \geq 0$ prior. Each of the three columns contains four rows of panels representing examples of learning paths that commonly occur for the given values of a and b. The starting point ($t = 1$) is indicated with a circle; its x coordinate is always $(S/W)_1 = 0.9$.

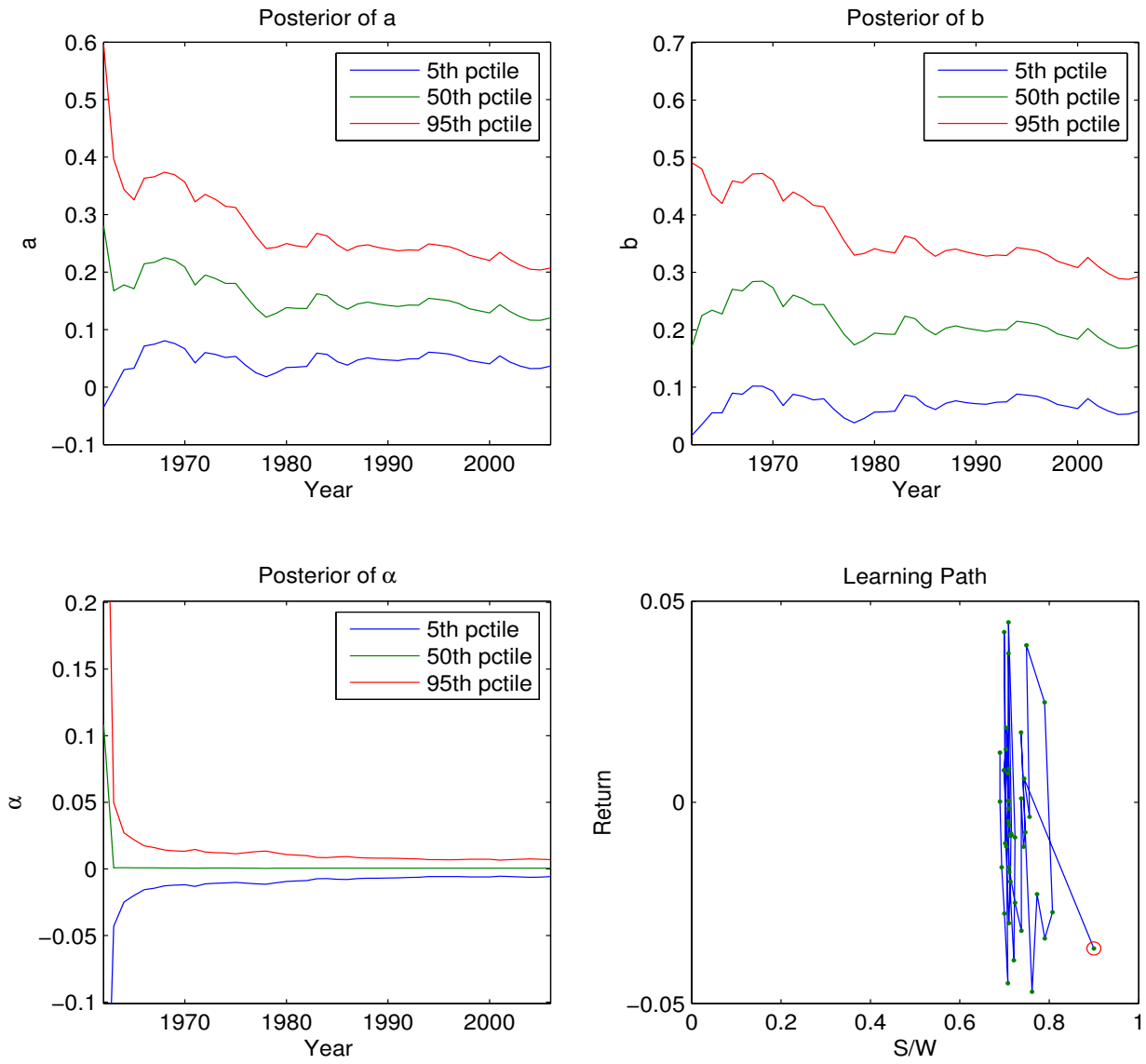

Figure 7. Results based on the year-by-year series of actual mutual fund returns under decreasing returns to scale. The first three panels plot the time series from 1962 through 2006 of selected percentiles of the posterior distributions of a, b, and $\alpha = a - b(S/W)$. The prior distribution is the $b \geq 0$ prior displayed in Figure 2. The data consist of actual annual returns on the aggregate portfolio of U.S. actively managed mutual funds, adjusted for exposures to the three Fama-French factors. The bottom right panel plots the sequence of equilibrium allocations and actual active fund returns. The starting allocation and first-year return are designated by the small circle.

References

Avramov, Doron, and Russ Wermers, 2006, Investing in mutual funds when returns are predictable, *Journal of Financial Economics* 81, 339–77.

Baks, Klaas P., Andrew Metrick, and Jessica Wachter, 2001, Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation, *Journal of Finance* 56, 45–85.

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

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

Chordia, Tarun, 1996, The structure of mutual fund charges, *Journal of Financial Economics* 41, 3–39.

Cuoco, Domenico, and Ron Kaniel, 2007, Equilibrium prices in the presence of delegated portfolio management, Working paper, Wharton and Fuqua.

Dangl, Thomas, Youchang Wu, and Josef Zechner, 2008, Market discipline and internal governance in the mutual fund industry, *Review of Financial Studies* 21, 2307–2343.

Das, Sanjiv R., and Rangarajan K. Sundaram, 2002, Fee speech: Signaling, risk-sharing, and the impact of fee structures on investor welfare, *Review of Financial Studies* 15, 1465–1497.

Dasgupta, Amil, Andrea Prat, and Michela Verardo, 2008, The price impact of institutional herding, Working paper, London School of Economics.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 2007, Disagreement, tastes, and asset prices, *Journal of Financial Economics* 83, 667–689.

Fama, Eugene F., and Kenneth R. French, 2009, Luck versus skill in the cross section of mutual fund α estimates, working paper, University of Chicago and Dartmouth College.

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

Fung, William, David A. Hsieh, Narayan Y. Naik, and Tarun Ramadorai, 2008, Hedge funds: Performance, risk, and capital formation, *Journal of Finance* 63, 1777–1803.

Garcia, Diego, and Joel M. Vanden, 2009, Information acquisition and mutual funds, *Journal of Economic Theory* 144, 1965–1995.

Glode, Vincent, 2009, Why mutual funds "underperform," Working paper, Wharton.

Glode, Vincent, and Richard C. Green, 2010, Information spillovers and performance persistence for hedge funds, Working paper, Wharton.

Grossman, Sanford G., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.

Gruber, Martin J., 1996, Another puzzle: The growth in actively managed mutual funds, *Journal of Finance* 51, 783–810.

Guerrieri, Veronica, and Peter Kondor, 2009, Fund managers, career concerns, and asset price volatility, Working paper, University of Chicago.

He, Zhiguo, and Arvind Krishnamurthy, 2008, Intermediary asset pricing, Working paper, University of Chicago.

Huang, Jennifer, Kelsey D. Wei, and Hong Yan, 2007, Participation costs and the sensitivity of fund flows to past performance, *Journal of Finance 62*, 1273–1311.

Investment Company Institute, 2009, *2009 Investment Company Fact Book*.

Jensen, Michael C., 1968, The performance of mutual funds in the period 1945–1964, *Journal of Finance* 23, 389–416.

Khorana, Ajay, Henri Servaes, and Peter Tufano, 2005, Explaining the size of the mutual fund industry around the world, *Journal of Financial Economics* 78, 145–185.

Kogan, Leonid, Stephen A. Ross, Jiang Wang, and Mark M. Westerfield, 2006, The price impact and survival of irrational traders, *Journal of Finance* 61, 195–229.

Lynch, Anthony W. and David K. Musto, 2003, How investors interpret past returns, *Journal of Finance* 58, 2033–2058.

Malkiel, Burton G., 1995, Returns from investing in equity mutual funds 1971 to 1991, *Journal of Finance* 50, 549–572.

Mamaysky, Harry, and Matthew Spiegel, 2002, A theory of mutual funds: Optimal fund objectives and industry organization, Working paper, Yale University.

Muthen, Bengt, 1990, Moments of the censored and truncated bivariate normal distribution, *British Journal of Mathematical and Statistical Psychology* 43, 131–143.

Nanda, Vikram, M.P. Narayanan, and Vincent A. Warther, 2000, Liquidity, investment ability, and mutual fund structure, *Journal of Financial Economics* 57, 417–443.

Pástor, Ľuboš, and Robert F. Stambaugh, 2002a, Mutual fund performance and seemingly unrelated assets, *Journal of Financial Economics* 63, 315–349.

Pástor, Ľuboš, and Robert F. Stambaugh, 2002b, Investing in equity mutual funds, *Journal of Financial Economics* 63, 351–380.

Petajisto, Antti, 2009, Why do demand curves for stocks slope down?, *Journal of Financial and Quantitiative Analysis* 44, 10131044.

Pollet, Joshua, and Mungo Wilson, 2008, How does size affect mutual fund behavior?, *Journal of Finance* 63, 2941–2969.

Rosenbaum, S., 1961, Moments of a truncated bivariate normal distribution, *Journal of the Royal Statistical Society, Series B (Methodological)* 21, 405–408.

Savov, Alexi, 2009, Free for a fee: The hidden cost of index fund investing, Working paper, University of Chicago.

Stein, Jeremy C., 2005, Why are most funds open-end? Competition and the limits of arbitrage, *Quarterly Journal of Economics* 120, 247–272.

Treynor, Jack L., and Fischer Black, 1973, How to use security analysis to improve portfolio selection, *Journal of Business* 46, 66–86.

Vayanos, Dimitri, and Paul Woolley, 2008, An institutional theory of momentum and reversal, Working paper, London School of Economics.

Wermers, Russ, 2000, Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses, *Journal of Finance* 55, 1655–1695.