On the Size of the Active Management Industry

by*

Ľuboš Pástor

and

Robert F. Stambaugh

January 5, 2010

Abstract

We analyze the equilibrium size of the active management industry and the role of historical data—how investors use it to decide how much to invest in the industry, and how researchers use it to judge whether the industry's size is reasonable. As the industry's size increases, every manager's ability to outperform passive benchmarks declines, to an unknown degree. We find that researchers need not be puzzled by the industry's substantial size despite the industry's negative track record. 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 Gene Fama, Vincent Glode, Ralph Koijen, Luke Taylor, Rob Vishny, Guofu Zhou, and workshop participants at Michigan State University, Ohio State University, and University of Chicago. Support as an Initiative for Global Markets Visiting Fellow (Stambaugh) at the University of Chicago is gratefully acknowledged.

1. Introduction

Active portfolio management remains popular, even though its overall 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 net returns 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]

We analyze the size of the active management industry in an equilibrium setting. Of particular interest is the role of historical data—how rational investors use it in theory and how researchers use it in practice. Rational investors use historical data to learn about how much they should invest in active management. Researchers use historical data to assess whether actual investor behavior is reasonable. We use our model in the first context to discover interesting endogeneity in the process by which investors learn. We use our model in the second context to ask whether researchers should be puzzled by the size of the active management industry.

We find that researchers need not be puzzled by the fact that active management remains popular despite its negative track record. Key to this conclusion is to realize that there are decreasing returns to scale in the active management industry—any manager's ability to outperform a benchmark declines as the industry's size increases. In contrast, active management's popularity would seem puzzling under the assumption of constant returns to scale, wherein a manager's ability to outperform would be the same at any size of the active management industry. The reason why active management's popularity is puzzling under constant returns but not under decreasing returns is that the industry's track record delivers very different messages under those two scenarios.

Under decreasing returns to scale, investors in our model learn about the degree of these decreasing returns and thereby determine the industry's equilibrium size. Researchers ask whether the industry's actual size is reasonable, given the unimpressive track record of the industry's returns. That track record leaves researchers quite uncertain about how much historical active returns would have improved had investors allocated less to active management. Given this uncertainty, researchers have a fairly wide confidence region for the active allocation that the investors in the model would currently choose. That confidence region includes active allocations that are sub-

[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 significantly 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)).

stantial. For example, we show that the active allocation can exceed 70% of investable wealth even if the industry's historical alpha is significantly negative. 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.

Under constant returns to scale, the industry's track record would lead investors to perceive a negative net alpha at any size for the industry, even if their prior beliefs about alpha were more optimistic than those leading to the results mentioned above under decreasing returns to scale. With the negative alpha, any positive investment in active management would be undesirable for mean-variance investors; they would instead go short if they could. Most studies that estimate mutual fund performance treat alphas as constant, unrelated to the industry's size. Under that scenario of constant returns to scale, the negative average alphas such studies typically report would make the industry's popularity puzzling, unlike under decreasing returns to scale.

Investors in our model face endogeneity that limits their learning about returns to scale in the active management industry. As they update their beliefs about the parameters governing returns to scale, they adjust the fraction of their investable wealth allocated to active management. They learn 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 the equilibrium allocation ceases to change from one period to the next, learning about returns to scale essentially stops. Interestingly, we find this is usually the case. The allocation converges to the level producing an alpha for the industry that appropriately compensates investors for non-benchmark risk. Investors eventually learn the alpha at that allocation, but they do not accurately learn the degree of decreasing returns to scale, even after thousands of years. Convergence of the allocation occurs quickly, after just a few years, when active returns are steeply decreasing in the industry's scale. When that relation is flatter, though, the industry's size can fluctuate at suboptimal levels for a long time before converging.

It seems reasonable to believe that a fund manager's ability to outperform a benchmark is decreasing in the aggregate amount of active management. As more money chases opportunities to outperform, prices are impacted and such opportunities become more elusive. If the benchmarks are sufficient for pricing assets in an efficient market, outperformance of the benchmarks 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.

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. Berk and Green set alphas to zero, whereas the alphas perceived by investors in our model are generally positive. Our model features competition among utility-maximizing investors and fee-maximizing fund managers, and the implications for alpha are derived in equilibrium. The equilibrium alpha is positive 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 industry. 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 there is no parameter uncertainty and non-benchmark risk can be completely diversified away. The net alpha investors receive in that case is zero whether or not managers compete, but the industry is significantly larger under competition. With many competing managers, managers 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 (as in our model) decreasing returns are such that the expected active return each manager produces declines linearly in the aggregate amount of active management. 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.

[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).

The paper is organized as follows. Section 2 describes the general setting of our model. Section 3 explores the model's equilibrium implications for alpha, fees, and the size of the active management industry in the absence of parameter uncertainty. The effects of parameter uncertainty are explored in Section 4. Section 5 discusses learning about returns to scale. Section 6 conducts inference about the size of the active management industry conditional on the industry's historical performance. Section 7 relates our model to that of Berk and Green (2004). Section 8 concludes.

2. Fund managers and investors: General 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 benchmarks. 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

$$r_F = \underline{\alpha} + Br_B + u, \tag{1}$$

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

$$r \;\; = \;\; \underline{\alpha} + u. \tag{2}$$

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

$$u_i = x + \epsilon_i, \tag{3}$$

for $i = 1, \ldots, M$, where all ϵ_i's have a mean of zero, variance of σ_ϵ^2, and zero correlation with

[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).

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 (3) 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 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

$$\pi_i = s_i \left(a - b \frac{S}{W} \right), \tag{4}$$

where s_i is the size of manager i's fund, with $S = \sum_{i=1}^{M} s_i$. The values of a and b can be either known or unknown, but we assume investors know that the values are identical across managers.

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 $a < 0$ in the setting in which a is unknown. 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

[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.

[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.

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 (4) 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 (4).

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,

$$\max_{f_i} f_i s_i. \tag{5}$$

Combining this fee structure with (4), we obtain the following relation for the ith element of $\underline{\alpha}$:

$$\alpha_i = a - b\frac{S}{W} - f_i. \tag{6}$$

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 $\gamma > 0$ 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. If the allocations to the benchmarks and riskless asset are unrestricted, then for each investor j the allocations to the funds solve the problem

$$\max_{\delta_j} \left\{ \delta_j' \mathrm{E}(r|D) - \frac{\gamma}{2} \delta_j' \mathrm{Var}(r|D) \delta_j \right\}, \tag{7}$$

where D denotes the set of information available to investors. We impose the restriction that the elements of the $M \times 1$ vector δ_j are non-negative (no shorting of funds). The next section analyzes the model in its simplest setting, in which a and b are assumed to be known. Subsequent sections then explore a setting where a and b remain uncertain after conditioning on D.

3. Equilibrium with a known profit function

In this section we explore the model when a and b are known, i.e. when D is sufficient to infer exactly the expected profit function in (4). We assume in this case that $a > 0$. Otherwise the non-negativity restriction on the elements of δ_j binds and there is no investment in the funds.

We solve for a symmetric Nash equilibrium among investors, wherein each investor solving (7) 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 fees in a symmetric Nash equilibrium among managers, who are solving (5). In equilibrium, all managers have the same alpha and charge the same fee, and all investors spread their wealth equally across all funds. That is, the M elements of δ are all identical, and the fraction of total wealth invested in active management, S/W, is given by the sum of those M elements. The following proposition gives the equilibrium values of the key quantities in the model.

Proposition 1. *In equilibrium for investors and managers when the values of a and b are known, we have $\alpha_i = \alpha$ and $f_i = f$ for $i = 1, \ldots, M$, where*

$$f = \frac{a\gamma\sigma_\epsilon^2}{2\gamma\sigma_\epsilon^2 + (M-1)p} \tag{8}$$

$$\alpha = a\left(1 - \frac{\gamma\sigma_\epsilon^2}{2\gamma\sigma_\epsilon^2 + (M-1)p}\right)\left(1 - \frac{Mb}{\gamma\sigma_\epsilon^2 + Mp}\right) \tag{9}$$

$$\frac{S}{W} = \frac{Ma}{\gamma\sigma_\epsilon^2 + Mp}\left(1 - \frac{\gamma\sigma_\epsilon^2}{2\gamma\sigma_\epsilon^2 + (M-1)p}\right), \tag{10}$$

where

$$p = \frac{N+1}{N}b + \gamma\sigma_x^2. \tag{11}$$

Proof: See Appendix.

All three quantities on the left-hand sides of (8) through (10) are positive. These quantities are analyzed in more detail in the following three subsections.

3.1. Fees

Equation (8) shows that the equilibrium fee f decreases in the number of managers, M, due to competition among managers. In the limit, the fees disappear: $f \to 0$ as $M \to \infty$. Note that f is the portion of the 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, the supplier and manager are price takers, and such discretionary quantities vanish. That doesn't mean that that the supplier sets a zero price or that the manager works for nothing. Any competitive proportional fee, which isn't under the 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.

Equation (8) also shows that the highest possible fee obtains for $M = 1$, in which case the single manager sets the monopolistic fee $f = a/2$. In general, the equilibrium fee f increases with a. For $M > 1$, f also increases with σ_ϵ and N, and it decreases with both σ_x and b.

3.2. Alphas

To obtain some insight into equilibrium alphas, consider a scenario with many funds, $M \to \infty$. In this limiting case, equation (9) simplifies into

$$\alpha = a\left(\frac{(1/N)b + \gamma\sigma_x^2}{[(N+1)/N]b + \gamma\sigma_x^2}\right). \tag{12}$$

Alpha in equation (12) increases with $\gamma\sigma_x^2$ and decreases with b and N. It does not depend on σ_ϵ because such risk can be fully diversified away across managers (unlike when M is finite).

Equation (12) helps us understand the interesting role that the number of investors, N, plays in determining fund alphas. In the limiting case $N \to \infty$, equation (12) simplifies into

$$\alpha = a\left(\frac{\gamma\sigma_x^2}{b + \gamma\sigma_x^2}\right). \tag{13}$$

In this case, $\alpha > 0$ only because investors demand compensation for residual risk. If this risk is completely diversifiable ($\sigma_x^2 \to 0$), then $\alpha \to 0$. In contrast, when N is finite, α remains positive even if $\sigma_x^2 \to 0$, as long as $b > 0$. Specifically, when $\sigma_x^2 \to 0$, α in equation (12) simplifies into

$$\alpha = \frac{a}{N+1}. \tag{14}$$

Note that α decreases in N. Alphas become smaller with more investors because each additional investor imposes a negative externality on the existing investors by diluting their returns. The 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. This externality also explains the above-mentioned positive relation between f and N. When N increases, the aggregate active investment increases, reducing the total profit earned by investors and managers. To induce less investment, the managers raise their fees.

A scenario with fewer funds brings into play two additional effects that work in opposite directions. On the one hand, a lower M results in higher fees, which push alphas down. On the other hand, a lower M requires higher alphas to compensate risk-averse investors for σ_ϵ. The net effect can go either way, depending on the magnitudes of the other quantities entering equation (9).

3.3. Size of the active management industry

When the number of investors is large, the size of the active management industry is governed by a familiar mean-variance result. 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 (7) is

$$\max_{S/W} \left\{ \left(\frac{S}{W}\right) E(r_A|D) - \frac{\gamma}{2}\left(\frac{S}{W}\right)^2 \text{Var}(r_A|D) \right\}. \tag{15}$$

The solution to this problem is given by

$$\frac{S}{W} = \frac{E(r_A|D)}{\gamma \text{Var}(r_A|D)}. \tag{16}$$

It is readily seen that the relation in (16) prevails in equilibrium as N grows large. When M is large as well, the size of the active management industry relative to investable wealth approaches

$$\frac{S}{W} = \frac{a}{b + \gamma\sigma_x^2}, \tag{17}$$

which is the limit of (10) as $M \to \infty$ and $N \to \infty$. The size of the industry therefore increases with a and decreases with b and $\gamma\sigma_x^2$, which is intuitive. Combining (17) with (13) gives

$$\frac{S}{W} = \frac{\alpha}{\gamma\sigma_x^2}. \tag{18}$$

Equations (16) and (18) coincide because $E(r_A|D) = \alpha$ and $\text{Var}(r_A|D) = \sigma_x^2$. These relations follow from equations (2) and (3) and the fact that all elements of δ are identical:

$$r_A \;=\; \frac{1}{M}\sum_{i=1}^{M} r_i \;=\; \alpha + x + \frac{1}{M}\sum_{i=1}^{M}\epsilon_i. \tag{19}$$

The mean of r_A in equation (19) is α. In this case with many funds, the variance of r_A is σ_x^2 because when $M \to \infty$, the variance of the last term in (19) goes to zero.

With fewer funds, diversifiable risk also plays a role, but the relation in (16) still holds. For example, with $N \to \infty$ but $M = 1$, it is readily verified using (9) and (10) that

$$\frac{S}{W} = \frac{\alpha}{\gamma(\sigma_x^2 + \sigma_\epsilon^2)}, \tag{20}$$

which again conforms to (16), noting from (19) that in this case $\text{Var}(r_A|D) = \sigma_x^2 + \sigma_\epsilon^2$. In general, it can be shown that the equilibrium value of S/W in equation (10) is smaller than or equal to the mean-variance solution in equation (16). The equality between (10) and (16) occurs only if

$b/N \to 0$, which is the case in the above examples with $N \to \infty$. When N is finite, (10) is smaller than (16) because investors internalize some of the externality discussed earlier.

The equilibrium size of the active management industry can also be measured relative to the size that maximizes expected total profit. Using equation (4), expected total profit is

$$\Pi = \sum_{i=1}^{M} \pi_i = S\left(a - b\frac{S}{W}\right), \tag{21}$$

which is maximized at

$$\frac{S^*}{W} = \frac{a}{2b}. \tag{22}$$

Combining (22) with (10) and (11) gives

$$\frac{S}{S^*} = 2\left(\frac{Mb}{M\left(\frac{N+1}{N}b + \gamma\sigma_x^2\right) + \gamma\sigma_\epsilon^2}\right)\left(\frac{(M-1)\left(\frac{N+1}{N}b + \gamma\sigma_x^2\right) + \gamma\sigma_\epsilon^2}{(M-1)\left(\frac{N+1}{N}b + \gamma\sigma_x^2\right) + 2\gamma\sigma_\epsilon^2}\right) \leq 2. \tag{23}$$

When $M = 1$,

$$\frac{S}{S^*} = \frac{b}{\frac{N+1}{N}b + \gamma(\sigma_x^2 + \sigma_\epsilon^2)} \leq 1. \tag{24}$$

A single manager is underinvested relative to the profit-maximizing size S^* unless $\sigma_x^2 + \sigma_\epsilon^2 \to 0$ and $N \to \infty$. One reason behind this underinvestment is the fee charged by the manager–monopolist. The underinvestment also reflects risk aversion of investors, who care not only about expected profits but also about the associated risk. If N is small, the underinvestment also reflects the fact that investors internalize some of the effect of decreasing returns to scale.

When $M \to \infty$,

$$\frac{S}{S^*} = \frac{2b}{\frac{N+1}{N}b + \gamma\sigma_x^2}, \tag{25}$$

so there can be underinvestment ($S < S^*$) or overinvestment ($S > S^*$). Overinvestment occurs when $\gamma\sigma_x^2$ is sufficiently small. One reason is that when managers reduce their fees, they do not fully internalize the reduction in expected profit that occurs when the lower fees induce higher investment. In the special case when there is no risk, $\sigma_x^2 \to 0$, equation (25) simplifies into

$$\frac{S}{S^*} = \frac{2N}{N+1}. \tag{26}$$

Full investment obtains only for $N = 1$; otherwise there is overinvestment. Investors invest more than the profit-maximizing amount because they do not fully internalize the reduction in profits caused by the greater amount invested. When $N \to \infty$, $S/S^* \to 2$, so that $S \to \bar{S}$, where $\bar{S} = aW/b$ is the size that equates the expected profit in (21) to zero (see Figure 1). As discussed

earlier, the equilibrium α in equation (14) goes to zero as $N \to \infty$. That is, the many investors invest up to the point at which all expected profit has been eliminated.

This special case with $S = \bar{S}$ and $\alpha = 0$ warrants a note. Even though the active management industry then provides no superior returns to investors, it can 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. When the industry's size is $\bar{S}$, its expected future profit is zero because its actions have eliminated some of that mispricing. By moving prices toward fair values, the industry provides a positive externality to asset markets.

In the maximization in (7), 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

$$\sum_{i=1}^{M} \delta_{i,j} \leq \bar{\delta}, \tag{27}$$

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 (27) states that the fraction of each investor's wealth placed in actively managed funds is at most $\bar{\delta}$. When (27) binds, S/W in equation (10) exceeds $\bar{\delta}$, and the equilibrium value of S/W instead equals $\bar{\delta}$. Also, as in the earlier unconstrained setting, $f \to 0$ as $M \to \infty$: perfect competition among managers drives the discretionary portion of the fee to zero even when the constraint (27) binds. When the constraint binds, however, alpha exceeds the level consistent with the mean-variance relation in (18) that otherwise obtains under perfect competition among managers and investors (i.e., with infinite M and N). That is,

$$\alpha > \gamma \sigma_x^2 \frac{S}{W}, \tag{28}$$

where $\alpha = a - \bar{\delta} b$. The Appendix includes a treatment of the case where (27) binds.

4. Uncertainty about returns to scale

We now analyze the model when the parameters a and b in equation (4) are unknown. We denote the expectation and the covariance matrix of a and b conditional on the available data by

$$\mathrm{E}\left(\begin{bmatrix} a \\ b \end{bmatrix} \mid D \right) = \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} \tag{29}$$

$$\mathrm{Var}\left(\begin{bmatrix} a \\ b \end{bmatrix} \mid D \right) = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}. \tag{30}$$

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 the mean-variance objective in (7). 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 $\tilde{a} \leq 0$, then $S/W = 0$. If $\tilde{a} > 0$, then S/W is given by the (unique) real positive solution to the equation*

$$0 = \tilde{a} - \frac{S}{W}\left[\tilde{b} + \gamma(\sigma_a^2 + \sigma_x^2)\right] + \left(\frac{S}{W}\right)^2 2\gamma\sigma_{ab} - \left(\frac{S}{W}\right)^3 \gamma\sigma_b^2. \tag{31}$$

If investors also face the constraint in (27) and the solution to (31) exceeds $\bar{\delta}$, then $S/W = \bar{\delta}$. Proof: See Appendix.

When the equilibrium value of S/W lies between 0 and 1, it obeys the same mean-variance relation in (16) as before. To see this, first note that given the equilibrium value of S/W, the benchmark-adjusted aggregate active fund return from equation (19) is given by

$$r_A = a - b\frac{S}{W} + x, \tag{32}$$

using (6) and the fact that the last term in (19) vanishes as $M \to \infty$. It follows from (32) that

$$\mathrm{E}(r_A|D) = \tilde{a} - \tilde{b}\frac{S}{W} \tag{33}$$

and

$$\mathrm{Var}(r_A|D) = \sigma_a^2 + \sigma_x^2 - 2\left(\frac{S}{W}\right)\sigma_{ab} + \left(\frac{S}{W}\right)^2 \sigma_b^2. \tag{34}$$

Equation (31) can then be rewritten in the image of the mean-variance relation in (16):

$$\begin{aligned} \frac{S}{W} &= \frac{\tilde{a} - \tilde{b}(S/W)}{\gamma\left[\sigma_a^2 + \sigma_x^2 - 2(S/W)\sigma_{ab} + (S/W)^2\sigma_b^2\right]} \\ &= \frac{\mathrm{E}(r_A|D)}{\gamma\mathrm{Var}(r_A|D)}, \end{aligned} \tag{35}$$

where the second equality uses (33) and (34). Also note that equation (33) represents the perceived alpha of the active management industry, and that an alternative expression for equation (34) is $\mathrm{Var}(r_A|D) = \sigma_x^2 + \sigma_\alpha^2$, where σ_α represents uncertainty about the industry's alpha.

Our analysis of learning explores a simple setting in which the single-period model developed above is applied repeatedly in successive periods. We assume that investors' risk aversion is $\gamma = 2$.

We also specify the volatility of the aggregate active benchmark-adjusted return as $\sigma_x = 0.02$, or 2% per year. That 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 constructed as in Fama and French (1993), using data for the 1962–2006 period.[7]

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,

$$\left[\begin{array}{c} a \\ b \end{array} \right] \sim N\left(E_0, V_0\right) I(b \geq 0), \tag{36}$$

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

$$E_0 = \left[\begin{array}{c} E_0^a \\ E_0^b \end{array} \right], \qquad V_0 = \left[\begin{array}{cc} V_0^{aa} & V_0^{ab} \\ V_0^{ab} & V_0^{bb} \end{array} \right]. \tag{37}$$

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

[7] 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 and factors.

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 (35), 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. 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 6. The reason is that the two priors are updated very differently after observing the same evidence. This updating is described in the following section.

4.2. Updating beliefs and equilibrium allocations

To analyze the learning mechanism and the resulting posterior distributions, we simulate 300,000 samples of active management returns and optimal allocations to active management. For each sample, we randomly draw the values of a and b from their prior distribution and hold them constant throughout the sample. In each year t, beginning with $t = 1$, we perform three steps.

First, we have investors use Proposition 2 to solve for $(S/W)_t$, the equilibrium allocation to active management, given their current beliefs about a and b. We bound the allocations between 0 and 1, so any equilibrium values exceeding one are set equal to one and any equilibrium values smaller than zero are set equal to zero. For $t = 1$, investors with Prior 2 optimally choose $(S/W)_1 = 0.9$, as discussed earlier. Investors with Prior 1 optimally choose a larger initial allocation, $(S/W)_1 = 1$, since Prior 1 is more optimistic about α. In fact, Prior 1 is so optimistic that in the absence of an upper bound on S/W, investors would invest 378% of their wealth actively.

Second, we construct the benchmark-adjusted active management return following equation (32) as $r_{A,t} = a - b(S/W)_t + x_t$, where x_t is drawn randomly from the normal distribution with mean zero and variance σ_x^2. Note that under Prior 2 ($b \geq 0$), the return investors earn, $r_{A,t}$, is affected by their choice of $(S/W)_t$: the more they invest, the lower their subsequent return. In contrast, there is no such relation under Prior 1 ($b = 0$).

Third, we let investors update their beliefs about a and b in a Bayesian fashion. They do so by running a time-series regression of returns on the equilibrium allocations. After observing $r_{A,t}$ and $(S/W)_t$, the available data in D consist of $y_t = [r_{A,1} \ldots r_{A,t}]'$ and $z_t = [(S/W)_1 \ldots (S/W)_t]'$. Investors regress y_t on $-z_t$ and a constant; the regression's intercept is a and the slope is b (see equation (32)). Recall that investors' prior beliefs for a and b are given by the bivariate truncated normal distribution in equation (36), 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,

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

$$E = V^{-1} \left(V_0^{-1} E_0 + \frac{1}{\sigma_x^2} Z_t' y_t \right), \tag{39}$$

where $Z_t = [\ \iota_t \ \ -z_t\ ]$. The posterior distribution of a and b is bivariate truncated normal as in equation (36), except that E_0 and V_0 are replaced by E and V from equations (38) and (39).[8] 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 (29) and (30).[9] These moments are then used to choose the equilibrium allocation $(S/W)_{t+1}$ in the following year, and the learning process continues by repeating the same three steps in year $t + 1$.

5. Learning About Returns to Scale

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 realized returns and the true values of a and b. The probability distributions of possible values for those

[8] 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.

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

quantities thus give rise to distributions of the posterior standard deviations of a, b, and α. Figure 3 displays the evolution of these distributions across time periods. Each panel plots selected percentiles of the distribution of the given standard deviation across the 300,000 samples.

The three left panels of Figure 3 correspond to Prior 1 ($b = 0$). In this case, a and α coincide, which is why the top and bottom left panels of Figure 3 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 3 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 3). In contrast, the speed of learning about α is quite similar in these two cases (compare the bottom two panels in Figure 3). Despite being unable to learn a and b very well, investors are able to learn α about as easily as when they know $b = 0$ a priori.

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 5.1. 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 3 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 $\alpha = a$.

Further results on learning when $b \geq 0$ are plotted in Figure 4. The top panels plot the distributions of the differences between the perceived and true values, $\tilde{a} - a$ and $\tilde{b} - b$, across the 300,000 samples. These distributions shrink rapidly in the first couple of years, reflecting initial learning about a and b, but they quickly flatten out. In contrast, the distribution of $\tilde{\alpha} - \alpha$, plotted in the bottom left panel, continues shrinking at the $\sqrt{t}$ rate as learning about α carries over beyond the first few years. These results are consistent with the posterior standard deviations in Figure 3.

We define the "true" S/W as the value that obtains when a and b are known, as given in equation (17). The final panel of Figure 4 plots the distribution of the differences between the equilibrium $(S/W)_t$ and the true S/W. These differences continue shrinking over time well beyond the first few years, resembling the pattern for α rather than a and b. The 25th and 75th percentiles meet at zero, indicating that both the equilibrium and the true S/W's are at the corner solutions of zero or one for at least half of all samples. The difference between the 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. Investors seem to gradually converge to the true optimal allocation, although the convergence can be slow.

5.1. Endogeneity in Learning

The key message from Figures 3 and 4 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 5 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 from equation (17) so that the true value of S/W that would obtain under knowledge of the true parameters is $S/W = 0.5$. The (a, b) pair is then used to generate random samples of active returns, which are used to update Prior 2. Each of the three columns in Figure 5 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 $(S/W)_1 = 0.9$.

The intuition for why $(S/W)_t$ tends to stop changing so quickly comes across most clearly when b is high. Our discussion here focuses on the four right-most panels of Figure 5, in which b is high. The learning paths in these four panels look very similar, so one description fits them all. Since $(S/W)_1 = 0.9 > 0.5$, investors initially overinvest in active management, so their true expected return is negative (even though they subjectively expect a small positive return). The first realized return is typically around -18%. Upon observing such a 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, typically around 9%.[10] From this high return, investors infer they should invest more than 0.3. Their investment in year 3, $(S/W)_3$, is already close to the true value of 0.5. In all four panels, S/W "converges" to its true value after about 3 years, in that only small deviations from 0.5 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 $a/(b + \gamma\sigma_x^2)$ (equation (17)). When b is high, the true value is approximately equal to a/b.[11] This approximate relation can be visualized in Figure 1. When b is high, the equilibrium

[10] 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.

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

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 in year 3 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 3. 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 5, 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 3 exhibit a sharp initial drop.

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 from the right panels of Figure 5, 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 3 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 story is similar when b is at its prior median, as shown in the four middle panels of Figure 5. Investors learn a lot initially while S/W exhibits substantial variation, but the speed of learning slows down considerably after a decade or so when S/W settles down to a narrow range close to the true S/W. The datapoints exhibit more dispersion compared to the high b case, but the basic patterns are similar. Therefore, the intuition presented for high b is relevant for the median sample as well. To various degrees, this intuition fits most of our samples. It does not fit the samples for which b is low, however, as discussed next.

5.2. Departures from Optimal Industry Size

The left panels of Figure 5 contain the results when b is low. The learning paths in these left panels are quite different from those in the middle and right panels. The first major difference is that it generally takes much longer for S/W to settle in a narrow range, if it settles at all during the 300-year period analyzed here. For example, in the third 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 (32)). Since realized returns do not help investors much in finding the optimal investment level, S/W keeps wandering around.

Another unique feature of low b is that S/W often settles at a level that is substantially different from the true S/W. For example, in the second panel on the left, 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, the expected return 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 5. 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 can observe S/W settling down at a level different from 0.5, even after 300 years.

It would appear from Figure 5 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.[12] 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 the same true S/W, 0.9, as before. 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

[12] To see that convergence to a different value cannot occur, note that at any interior value to which S/W converges, (35) holds. After infinitely many realizations of returns at a given S/W, $\mathrm{E}(r_A|D) = \alpha$, the true alpha in equation (18) at that value of S/W. With no uncertainty about alpha, $\mathrm{Var}(r_A|D) = \sigma_x^2$, which appears in the denominator of (18). Therefore the value to which S/W converges must be the "true" value that satisfies (18) and thus (17).

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. For all practical purposes, we can 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 long periods of time.

Let us briefly summarize the key findings from Figure 5. 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, 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.

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

In this section, we take the perspective of a researcher who asks whether it is puzzling how large the active management industry is, given its poor historical performance. Conditional on that performance, the researcher forms a posterior distribution for the current equilibrium S/W based on our model. We show that this posterior crucially depends on the researcher's prior beliefs about b, the parameter governing returns to scale in the industry. The researcher's prior beliefs about a and b are assumed to be the same as the investors', for simplicity.

The researcher uses the posterior distribution for the current equilibrium S/W 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 the active funds, passive benchmarks, and the riskless asset. The investors' holdings of the riskless asset, or cash, 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.

In computing the posterior for S/W conditional on the 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. The t-statistic is computed as $t = \hat{\alpha}\sqrt{T}/\sigma_x$ for $T = 50$ years.[13] The posterior distribution is obtained from simulated samples generated as described earlier: for each sample, a and b are drawn from the prior, and then in each year of the sample a

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

return is drawn and the new S/W is computed. The posterior distribution for S/W conditional on a given value t_0 of the t-statistic is constructed as the distribution of the S/W values for year 50 in all samples producing t-statistics within a small neighborhood of t_0. Figure 6 plots the resulting posterior distributions for t_0 ranging from -4 to 4.

Panel A of Figure 6 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 then collapses to a single value because the t-statistic is a sufficient statistic for S/W under this prior. 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 at all in active management. This theoretical result, obtained under $b = 0$, 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 6 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 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 (13). 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.

To summarize, given a track record representative of active U.S. equity mutual funds, investors who believe that $b = 0$ would invest nothing in active management. However, it seems more reasonable to allow for decreasing returns to scale, or $b \geq 0$. Following the same track record, investors with such prior beliefs often find it optimal to invest a substantial fraction of their wealth in active management, even though their prior beliefs about α are more pessimistic than the beliefs of the $b = 0$ investors. In short, it is not a puzzle that active management remains popular, despite its track record.

6.1. Robustness

Figures 2 through 6 present results for two particular sets of prior beliefs about a and b. The prior for a in Figure 2 might seem optimistic, but optimism is not the driving force behind our results. After all, we have seen that this same prior leads investors with $b = 0$ to invest nothing in active management given its negative track record. Our results are instead driven by the prior on b. To support this statement, we also present results for an alternative set of priors, which are less optimistic than the priors plotted in Figure 2. 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 alternative prior for a is plotted in Figure 7. This prior assigns a 26% probability to the event that $a < 0$, which is substantially larger than the 7.2% probability in Figure 2. The implied prior for α is also more pessimistic, with the median ranging from 0 to 0.2 and the 5th percentile ranging from about -0.6 to -0.3.

The results for this alternative prior are plotted in Figure 8. Similar to Figure 6, under $b = 0$,

$S/W = 0$ when the track record is negative, whereas under $b \geq 0$, the equilibrium S/W is substantial. For example, conditional on $t = -2$, the median S/W is 0.10 and the 95th percentile is 0.65. These values are smaller than their counterparts in Figure 6, but that is not surprising: before seeing any data, investors allocate only half of their investable wealth to active management, so after seeing a negative track record, they generally allocate less than half. More important, the results based on this more pessimistic prior confirm that with decreasing returns to scale, active management can remain popular in equilibrium despite its negative track record.

7. 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.[14] 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. Assuming decreasing returns at the individual fund level seems plausible as well, though it encounters the question of 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 (4), 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.[15] As discussed earlier, when both a and b in (4) 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 $\alpha = 0$, whereas our investors perceive

[14] 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).

[15] 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 (1)).

$\alpha > 0$. We solve for the Nash equilibrium among investors maximizing (7). BG do not solve the investors' optimization problem explicitly; instead, they fix $\alpha = 0$ 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; $\alpha = 0$ is therefore a necessary condition for equilibrium. Recall from equations (13) and (14) that our model implies $\alpha \to 0$ as well if there are many funds ($M \to \infty$) and all non-benchmark risk is diversifiable ($\sigma_x \to 0$), as long as the number of investors is very large ($N \to \infty$). With a smaller number of investors, however, $\alpha > 0$ because investors internalize some of the reduction in alpha caused by their own investment.

As discussed earlier, it seems reasonable that non-benchmark risk cannot be fully diversified across actively managed funds, so that $\sigma_x > 0$. As a result, $\alpha > 0$ even with many funds and many investors. However, we do not wish to leave readers with the impression that alpha in that setting is necessarily large. Once learning proceeds to the point where uncertainty about alpha is small, the non-benchmark variance is essentially just σ_x^2, and alpha is then equal to $\gamma\sigma_x^2$ times the equilibrium allocation S/W, as implied by equation (18). Even with $S/W = 1$, the values of γ and σ_x specified in our numerical investigation (2 and 0.02, respectively) imply a value of α equal to only 8 basis points per annum.[16] 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 where $\sigma_x > 0$ and there are many funds and investors.[17]

In their diversification argument justifying $\alpha = 0$, 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 is S^* in our setting.[18] In our model, with many 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, BG's assumption that $\sigma_x = 0$ implies that the amount invested is $\bar{S} = 2S^*$. That is, the industry's size then reaches the level that produces zero expected profit.

[16] If α remains uncertain, this calculation is modified by replacing σ_x^2 with $\sigma_x^2 + \sigma_\alpha^2$, where σ_α^2 denotes the posterior variance of α (see equations (33) through (35)). Uncertainty about α thus increases the equilibrium value of α, but for realistic parameter values, this increase amounts to only a few basis points per annum. The value of α increases further when the number of investors is finite (equation (14)), but that effect is small unless N is very small.

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

[18] This profit-maximizing amount of active management for a given fund is denoted as q_t^* by BG. Their equation (26), $q_t^*(\phi_t) = \phi_t/2a$, corresponds directly to our equation (22). Their a corresponds to our b, their ϕ_t corresponds to (the expected value of) our a, and their q_t corresponds to our S/W.

The specification that brings our model closest to that of BG involves many investors ($N \to \infty$), a single manager ($M = 1$), and no risk ($\sigma_\epsilon \to 0$ and $\sigma_x \to 0$). With $M = 1$, we obtain $f = a/2$, as in BG.[19] With $N \to \infty$, the externality present with fewer investors, which is absent in BG, disappears. To obtain BG's condition that $\alpha = 0$ when there is a single fund ($M = 1$), that fund can have no risk; otherwise investors would require $\alpha > 0$. The absence of risk requires $\sigma_x \to 0$ as well as $\sigma_\epsilon \to 0$. With $M \to \infty$, σ_ϵ would drop out (this is BG's diversification argument), but since we need $M = 1$ to replicate BG's value for f, we also need $\sigma_\epsilon \to 0$ to obtain $\alpha = 0$. Finally, using (24), we see that equilibrium under the above specification produces $S = S^*$, the profit-maximizing size of the industry that is analogous to the profit-maximizing fund size obtained in BG.

8. 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 underperformance of active funds over the past few decades, our model predicts that the investors' proportional allocation to active management should have decreased over time. Indeed, passive indexing has grown dramatically since its beginnings in the 1970s, consistent with the model.

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.

[19] 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.

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 the relations given in Propositions 1 and 2. We first analyze the case in which the parameters a and b of the profit function are known.

A.1. Known Profit Function

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

$$s = \frac{W}{N} \sum_{n=1}^{N} \delta_n \tag{A1}$$

$$S = \iota_M' s = \frac{W}{N} \sum_{n=1}^{N} \iota_M' \delta_n, \tag{A2}$$

where ι_M is an $M \times 1$ vector of ones. We can express the first two moments of fund returns as

$$\mathrm{E}(r|D) = \underline{\alpha} = a\iota_M - \frac{b}{N} \sum_{n=1}^{N} \iota_M' \delta_n \iota_M - \underline{f} \tag{A3}$$

$$\mathrm{Var}(r|D) = \Omega, \tag{A4}$$

where $\underline{\alpha}$ is an $M \times 1$ vector whose i-th element is α_i, $\underline{f}$ is an $M \times 1$ vector whose i-th element is f_i, and Ω denotes the covariance matrix of u (since a and b are known), given by

$$\Omega = \sigma_x^2 \iota_M \iota_M' + \sigma_\epsilon^2 I_M. \tag{A5}$$

Above, I_M denotes the $M \times M$ identity matrix. Substituting these relations into (7) then gives investor j's problem as

$$\max_{\delta_j} \left\{ \delta_j' \left(a\iota_M - \frac{b}{N} \sum_{n \neq j} \iota_M' \delta_n \iota_M - \underline{f} \right) - \delta_j' \frac{b}{N} \iota_M \iota_M' \delta_j - \frac{\gamma}{2} \delta_j' \Omega \delta_j \right\}, \tag{A6}$$

subject to the restrictions

$$\iota_M' \delta_j \leq \bar{\delta} \tag{A7}$$

$$\delta_{i,j} \geq 0 \ \ \forall i, j, \tag{A8}$$

where $\delta_{i,j}$ denotes the i-th element of δ_j. We here impose the leverage constraint in (A7) from the outset and then simply obtain Proposition 1 as the case when it 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 is

$$a\iota_M - \underline{f} - \frac{b}{N} \iota_M \iota_M' (\sum_{n=1}^{N} \delta_n + \delta_j) - \gamma \Omega \delta_j - \lambda_1 \iota_M - \lambda_2 = 0, \tag{A9}$$

where the scalar λ_1 and the $M \times 1$ vector λ_2 contain the multipliers associated with the constraints in (A7) and (A8). Since all investors are identical, so are their equilibrium allocations across funds:

$$\delta_j = \delta, \qquad j = 1, \ldots, N \tag{A10}$$

Imposing (A10) on the first-order conditions then implies

$$a\iota_M - \underline{f} - \frac{(N+1)}{N} b\iota_M \iota_M' \delta - \gamma\Omega\delta - \lambda_1 \iota_M - \lambda_2 = 0, \tag{A11}$$

which yields δ as a function of $\underline{f}$:

$$\delta = G(a\iota_M - \underline{f} - \lambda_1 \iota_M - \lambda_2), \tag{A12}$$

where

$$G = \left(\frac{N+1}{N} b\iota_M \iota_M' + \gamma\Omega \right)^{-1}. \tag{A13}$$

The M managers, who understand the relation in (A12), set their fees before investors make their decisions. Each manager i chooses f_i to maximize

$$f_i s_i = f_i \frac{W}{N} \sum_{n=1}^{N} \delta_{i,n} = f_i \frac{W}{N} N \delta_{(i)} = f_i W \delta_{(i)}, \tag{A14}$$

where $\delta_{(i)}$ is a scalar denoting the amount that each investor invests in fund i. It follows from equation (A12) that

$$\delta_{(i)} = g_i'(a\iota_M - \underline{f} - \lambda_1 \iota_M - \lambda_2), \tag{A15}$$

where g_i' denotes the i-th row of G. Each manager i thus solves

$$\max_{f_i} \left\{ f_i g_i'(a\iota_M - \underline{f} - \lambda_1 \iota_M - \lambda_2) \right\}. \tag{A16}$$

The first-order condition, taking other managers' fees as given, is

$$g_i'(a\iota_M - \lambda_1 \iota_M - \lambda_2) - g_i'(\underline{f} + e_i f_i) = 0, \tag{A17}$$

where e_i is an $M \times 1$ vector whose i-th element is one and the other $M - 1$ elements are zero. Since all managers are identical, their equilibrium fees are the same:

$$f_i = f, \qquad i = 1, \ldots, M, \tag{A18}$$

so that $\underline{f} = f\iota_M$. As a result, all funds are held in equilibrium in nonnegative amounts, so the constraint (A8) does not bind and we can set $\lambda_2 = 0$ throughout. Substituting into the first-order condition, we obtain

$$g_i'(a\iota_M - \lambda_1 \iota_M) = g_i'(\iota_M + e_i) f. \tag{A19}$$

This equation must hold for all $i = 1, \ldots, M$, so that

$$G(a\iota_M - \lambda_1 \iota_M) = [G + \text{diag}(G)]\, \iota_M f, \tag{A20}$$

from which we obtain

$$f = [G + \text{diag}(G)]^{-1}\, G(a - \lambda_1). \tag{A21}$$

To invert the matrix, we make use of the easily verified relation,

$$(p_1 I_M + p_2 \iota_M \iota_M')^{-1} = \frac{1}{p_1} I_M - \frac{p_2}{p_1(p_1 + Mp_2)} \iota_M \iota_M'. \tag{A22}$$

Based on equations (A5) and (A13), we can write

$$\begin{aligned} G^{-1} &= \frac{N+1}{N} b \iota_M \iota_M' + \gamma \left(\sigma_x^2 \iota_M \iota_M' + \sigma_\epsilon^2 I_M \right) && \text{(A23)} \\ &= p_1 I_M + p_2 \iota_M \iota_M', && \text{(A24)} \end{aligned}$$

where

$$\begin{aligned} p_1 &= \gamma \sigma_\epsilon^2 && \text{(A25)} \\ p_2 &= \frac{N+1}{N} b + \gamma \sigma_x^2. && \text{(A26)} \end{aligned}$$

Using the relation in (A22), we obtain

$$G = \frac{1}{p_1} I_M - \frac{p_2}{p_1(p_1 + Mp_2)} \iota_M \iota_M'. \tag{A27}$$

Since

$$\text{diag}(G) = \left(\frac{1}{p_1} - \frac{p_2}{p_1(p_1 + Mp_2)} \right) I_M, \tag{A28}$$

we have

$$G + \text{diag}(G) = p_3 I_M + p_4 \iota_M \iota_M', \tag{A29}$$

where

$$\begin{aligned} p_3 &= \frac{2}{p_1} - \frac{p_2}{p_1(p_1 + Mp_2)} && \text{(A30)} \\ p_4 &= -\frac{p_2}{p_1(p_1 + Mp_2)}. && \text{(A31)} \end{aligned}$$

Invoking (A22) again, we have

$$[G + \text{diag}(G)]^{-1} = \frac{1}{p_3} I_M - \frac{p_4}{p_3(p_3 + Mp_4)} \iota_M \iota_M'. \tag{A32}$$

We also have

$$G\iota_M = \left(\frac{1}{p_1}I_M - \frac{p_2}{p_1(p_1+Mp_2)}\iota_M\iota_M'\right)\iota_M \tag{A33}$$

$$= \left(\frac{1}{p_1} - \frac{Mp_2}{p_1(p_1+Mp_2)}\right)\iota_M \tag{A34}$$

$$= \frac{1}{p_1+Mp_2}\iota_M. \tag{A35}$$

Combining equations (A32) and (A34), we have

$$[G+\text{diag}(G)]^{-1}G\iota_M = \frac{1}{p_1p_3}\left(I_M - \frac{p_4}{p_3+Mp_4}\iota_M\iota_M'\right)\left(1-\frac{Mp_2}{p_1+Mp_2}\right)\iota_M \tag{A36}$$

$$= \frac{1}{(p_1+Mp_2)(p_3+Mp_4)}\iota_M. \tag{A37}$$

From the definitions of p_3 and p_4 in (A30) and (A31), we have

$$p_3+Mp_4 = \frac{2}{p_1} - \frac{p_2}{p_1(p_1+Mp_2)} - \frac{Mp_2}{p_1(p_1+Mp_2)} \tag{A38}$$

$$= \frac{1}{p_1}\left(2-\frac{(M+1)p_2}{p_1+Mp_2}\right). \tag{A39}$$

Substituting this into (A37) and rearranging terms, we obtain

$$(G+\text{diag}(G))^{-1}G\iota_M = \frac{p_1}{2p_1+(M-1)p_2}\iota_M. \tag{A40}$$

Combining equations (A21) and (A40), we obtain

$$f = \frac{p_1(a-\lambda_1)}{2p_1+(M-1)p_2}. \tag{A41}$$

Substituting this into (A12), recalling that $\lambda_2 = 0$, gives

$$\begin{aligned}\delta &= G(a\iota_M - \frac{p_1(a-\lambda_1)}{2p_1+(M-1)p_2}\iota_M - \lambda_1\iota_M)\\ &= (a - \frac{p_1(a-\lambda_1)}{2p_1+(M-1)p_2} - \lambda_1)G\iota_M\\ &= \left(\frac{a-\lambda_1}{p_1+Mp_2}\right)\left(1-\frac{p_1}{2p_1+(M-1)p_2}\right)\iota_M,\end{aligned} \tag{A42}$$

where the last equality uses (A35).

When the constraint in (A7) does not bind, we can set $\lambda_1 = 0$ in (A41) and (A42) to obtain,

$$s = \delta W = \left(\frac{aW}{p_1+Mp_2}\right)\left(1-\frac{p_1}{2p_1+(M-1)p_2}\right)\iota_M \tag{A43}$$

$$\frac{S}{W} = \frac{\iota_M's}{W} = \left(\frac{aM}{p_1+Mp_2}\right)\left(1-\frac{p_1}{2p_1+(M-1)p_2}\right) \tag{A44}$$

$$f = \frac{ap_1}{2p_1+(M-1)p_2}, \tag{A45}$$

where the first two equations also invoke (A1) and (A2). Substituting (A43) through (A45) into (6) and rearranging then gives $\underline{\alpha} = \alpha\iota_M$, where

$$\alpha = a\left(1 - \frac{bM}{p_1 + Mp_2}\right)\left(1 - \frac{p_1}{2p_1 + (M-1)p_2}\right). \tag{A46}$$

Equations (8) through (10) in Proposition 1 follow immediately from (A44) through (A46). Also note that p defined in (11) is identical to p_2 in (A26).

When the constraint in (A7) binds, we can substitute δ from equation (A42) into the constraint $\bar{\delta} = \iota_M'\delta$, giving

$$\bar{\delta} = \iota_M'\delta = \frac{M(a-\lambda_1)}{p_1 + Mp_2}\left(1 - \frac{p_1}{2p_1 + (M-1)p_2}\right). \tag{A47}$$

Solving for λ_1 yields

$$\lambda_1 = a - \frac{\bar{\delta}(p_1 + Mp_2)}{M}\left(\frac{2p_1 + (M-1)p_2}{p_1 + (M-1)p_2}\right), \tag{A48}$$

which when substituted in (A41) gives, after simplifying,

$$f = \frac{\bar{\delta}p_1}{M}\frac{p_1 + Mp_2}{p_1 + (M-1)p_2}. \tag{A49}$$

Note also that $\delta = (1/M)\bar{\delta}\iota_M$ since $\iota_M'\delta = \bar{\delta}$, and thus

$$s = \delta W = \frac{\bar{\delta}W}{M}\iota_M \tag{A50}$$

$$\frac{S}{W} = \frac{\iota_M' s}{W} = \bar{\delta}. \tag{A51}$$

Substituting from (A49) through (A51) into (6) gives $\underline{\alpha} = \alpha\iota_M$, where

$$\alpha = a - \frac{\bar{\delta}}{M}\left[Mb + p_1\frac{p_1 + Mp_2}{p_1 + (M-1)p_2}\right]. \tag{A52}$$

For $M \to \infty$, $\alpha \to a - \bar{\delta}b$, and α satisfies the inequality in (28) that otherwise holds as an equality when S/W satisfies (10).

A.2. Unknown Profit Function

When a and b are unknown, with their conditional moments given in (29) and (30), the vector of benchmark-adjusted fund returns is given by

$$\begin{aligned} r &= a\iota_M - b\frac{S}{W}\iota_M - \underline{f} + u \qquad (A53)\\ &= \tilde{a}\iota_M - \tilde{b}\frac{S}{W}\iota_M - \underline{f} + u + (a-\tilde{a})\iota_M - (b-\tilde{b})\frac{S}{W}\iota_M \\ &= \tilde{a}\iota_M - \tilde{b}\frac{1}{N}\sum_{n=1}^{N}\iota_M'\delta_n\iota_M - \underline{f} + \left\{u + (a-\tilde{a})\iota_M - (b-\tilde{b})\frac{1}{N}\sum_{n=1}^{N}\iota_M'\delta_n\iota_M\right\}, \end{aligned}$$

so that

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

where

$$\sigma_1^2 = \sigma_x^2 + \sigma_a^2. \tag{A54}$$

When facing the problem in (7), 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

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

subject to the constraints (A7) and (A8). 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:

$$0 = M(\tilde{a}-f) - 2\delta_{(j)}\tilde{b}\frac{M^2}{N} - \tilde{b}\frac{M^2(N-1)}{N}\delta^*$$

$$-\gamma\delta_{(j)}M^2\sigma_1^2 - \gamma\delta_{(j)}M\sigma_\epsilon^2 + 3\gamma\delta_{(j)}^2\sigma_{ab}\frac{M^3}{N} + 2\gamma\delta_{(j)}\sigma_{ab}\frac{M^3(N-1)}{N}\delta^*$$
$$-2\gamma\delta_{(j)}^3\sigma_b^2\frac{M^4}{N^2} - 3\gamma\delta_{(j)}^2\sigma_b^2\frac{M^4(N-1)}{N^2}\delta^* - \gamma\delta_{(j)}\sigma_b^2\frac{M^4(N-1)^2}{N^2}(\delta^*)^2 - M\lambda_1 - \lambda_2.$$

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

$$\frac{S}{W} = \iota_M'\frac{s}{W} = \iota_M'\frac{1}{N}\sum_{n=1}^{N}\delta_n = \iota_M'\frac{1}{N}N\delta^*\iota_M = M\delta^*. \tag{A55}$$

As $M \to \infty$ and $N \to \infty$, the first-order condition then becomes, using (A55),

$$0 = \tilde{a} - f - \lambda_1 - \frac{S}{W}\left[\tilde{b} + \gamma\sigma_1^2\right] + \left(\frac{S}{W}\right)^2 2\gamma\sigma_{ab} - \left(\frac{S}{W}\right)^3\gamma\sigma_b^2. \tag{A56}$$

Following the earlier analysis, we set $f = 0$ when the number of funds (M) is infinite. When the constraint in (A8) does not bind and thus $\lambda_1 = 0$, (A56) is identical to equation (31) in Proposition 2, noting (A54). It can be verified that this equation has one positive real solution for S/W. If that solution exceeds $\bar{\delta}$, then (A50) and (A51) apply as before.

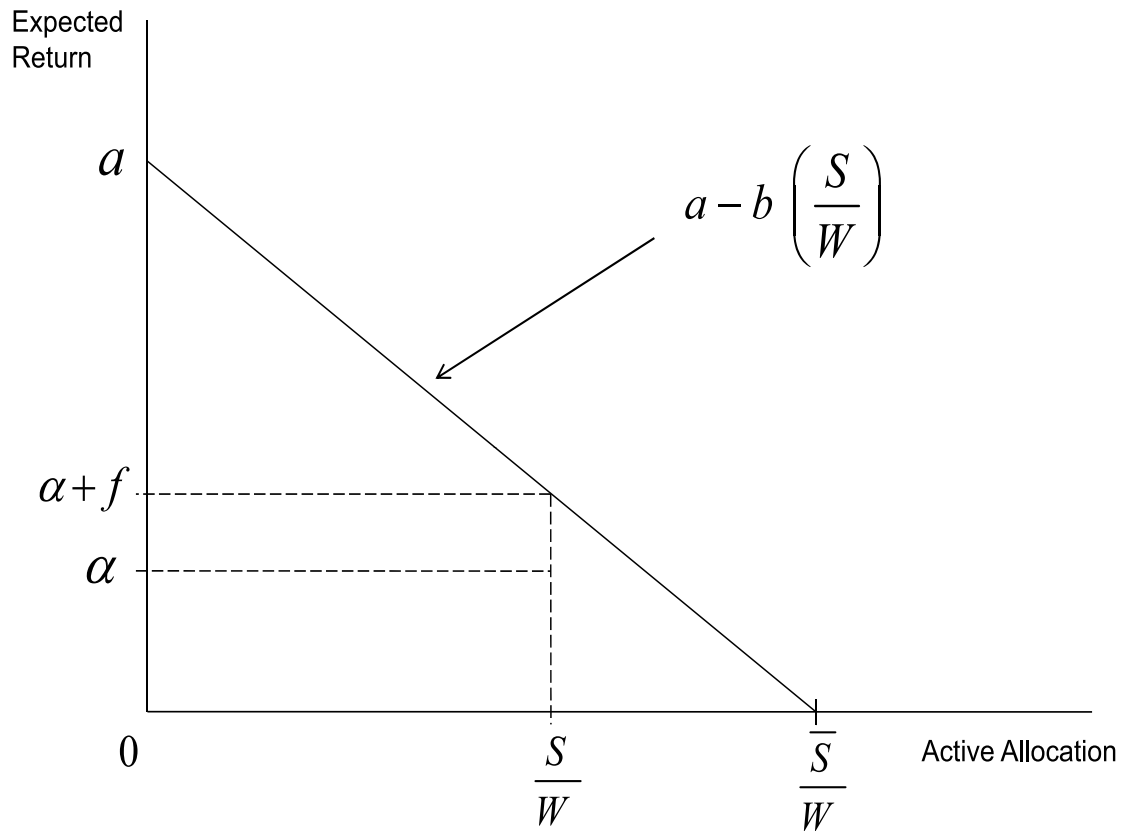

Figure 1. Decreasing returns to scale in the active management industry. This figure plots the theoretical relation between the expected benchmark-adjusted excess fund return before fees against the relative size of the active management industry. Specifically, it plots equation (6): $\alpha + f = a - b\frac{S}{W}$, 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 $\alpha = f = 0$.

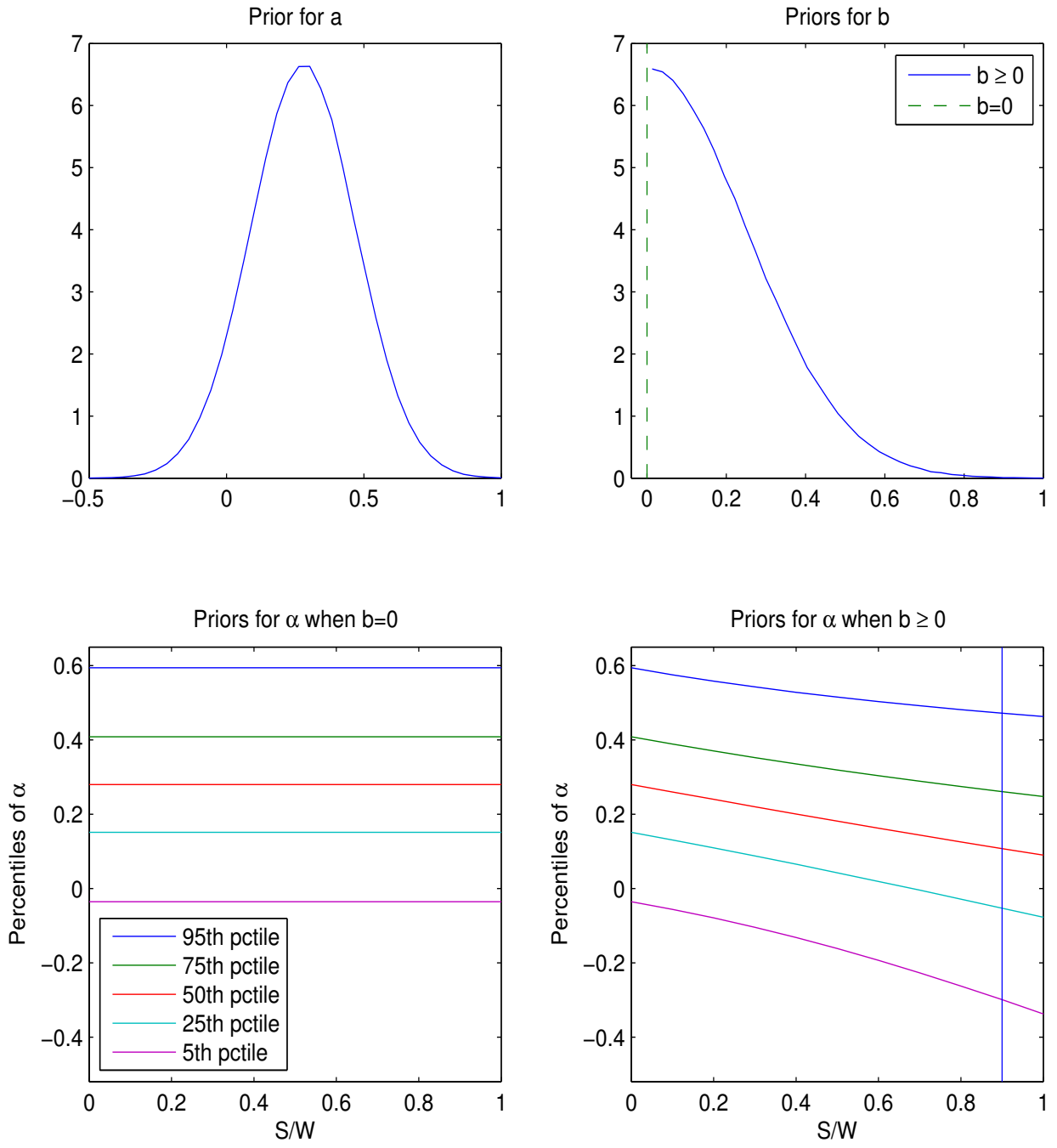

Figure 2. Prior distributions. This figure plots the prior distributions for the parameters of the function in equation (6). Panel A plots the prior for a, which is normal with the mean of 0.28 and standard deviation of 0.19. Panel B plots two different prior distributions for b: $b = 0$ (constant returns to scale, known b), and $b \geq 0$ (decreasing returns to scale, unknown b). 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.

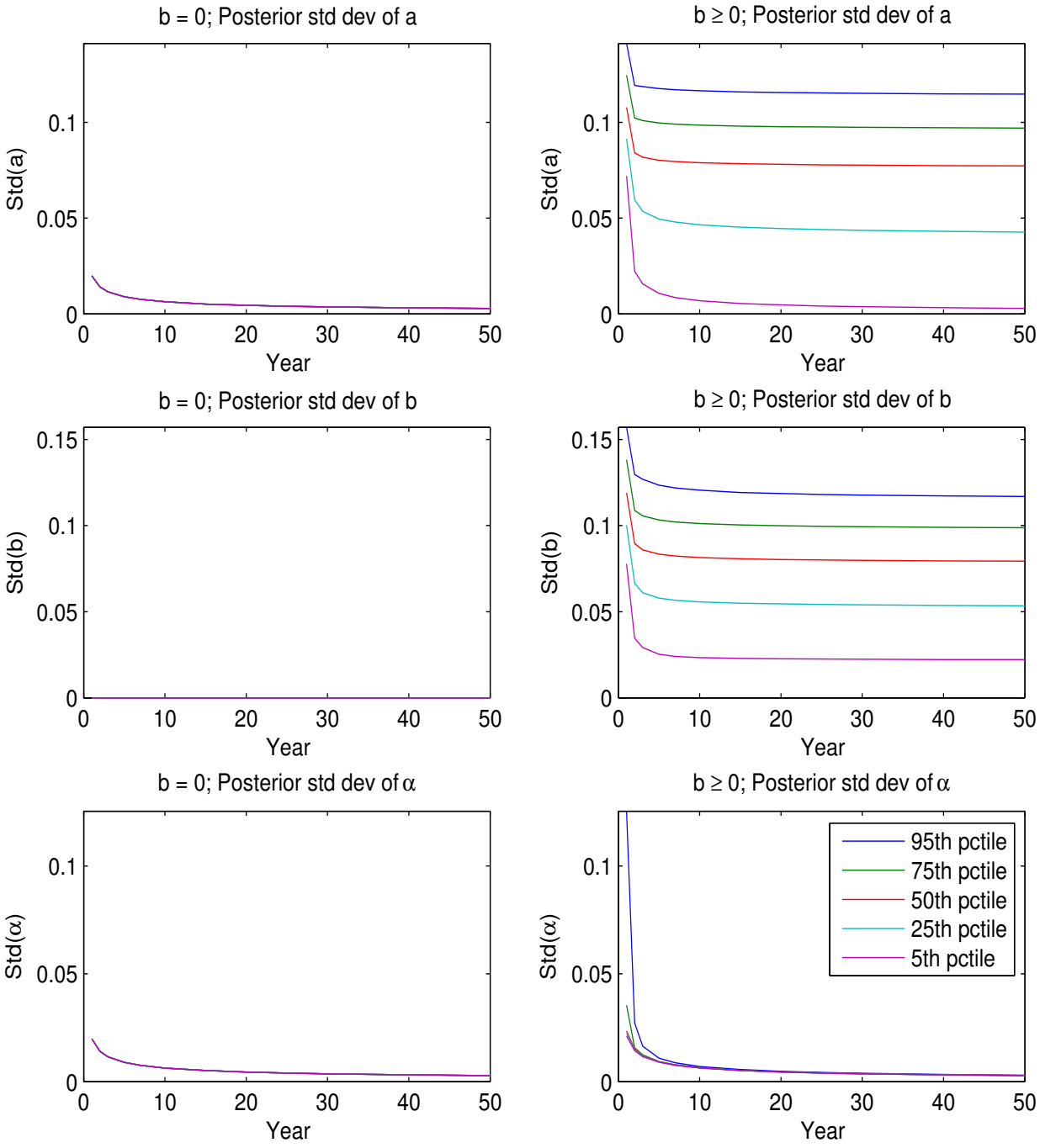

Figure 3. Posterior standard deviations. This figure plots the posterior standard deviations of a, b, and α as a function of time. The three panels on the left correspond to the prior $b = 0$; the three panels on the right represent the prior $b \geq 0$. Each panel on the right plots the 5th, 25th, 50th, 75th, and 95th 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 $b = 0$.

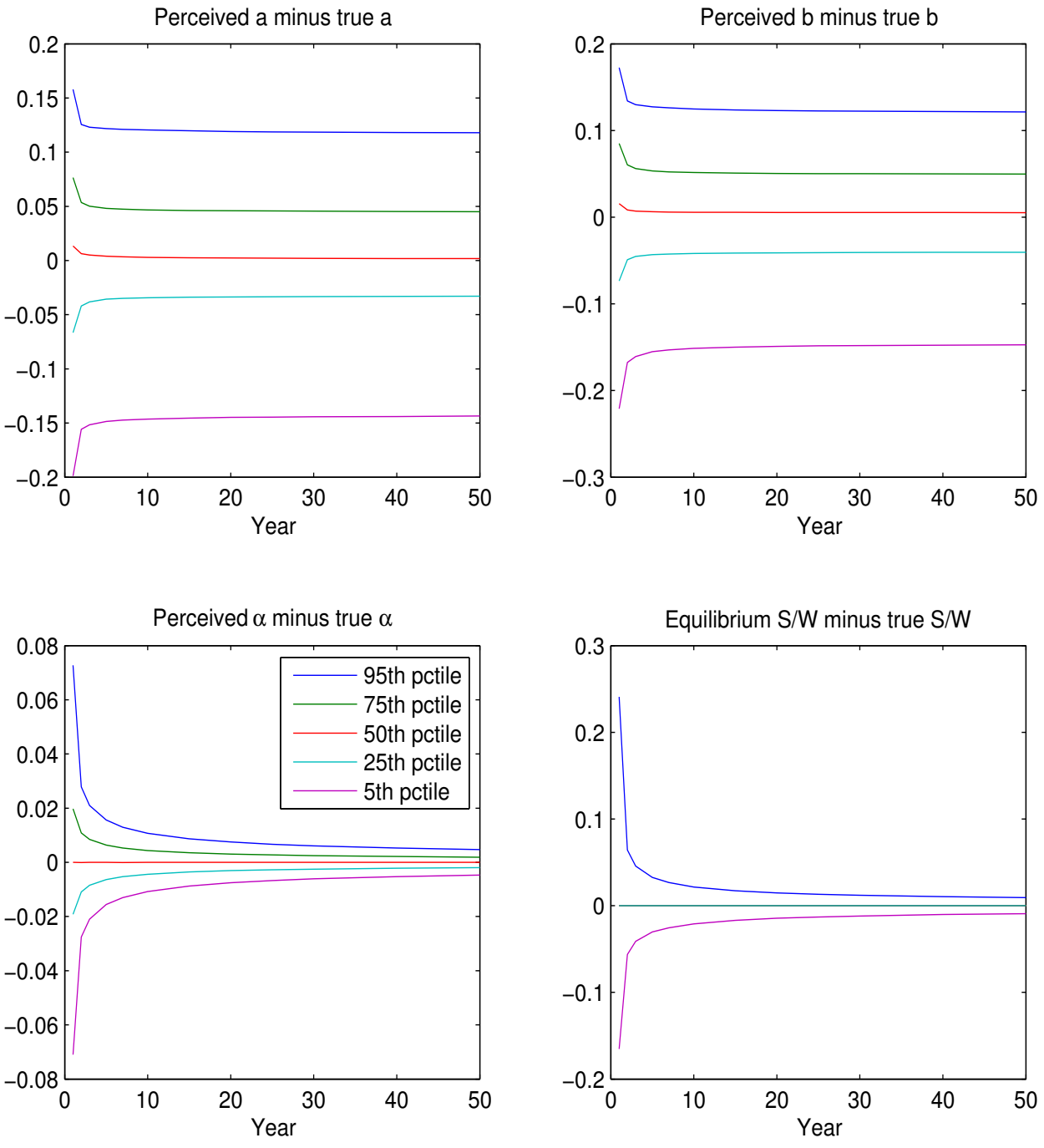

Figure 4. Deviations from true values. Panels A, B, and C plot the distributions of the differences between the perceived and true values, $\tilde{a} - a$, $\tilde{b} - b$, and $\tilde{\alpha} - \alpha$, respectively, across many simulated samples under the $b \geq 0$ prior. Panel D plots the distribution of the differences between the equilibrium $(S/W)_t$ and the "true" S/W, where the latter quantity is computed from the true values of a and b.

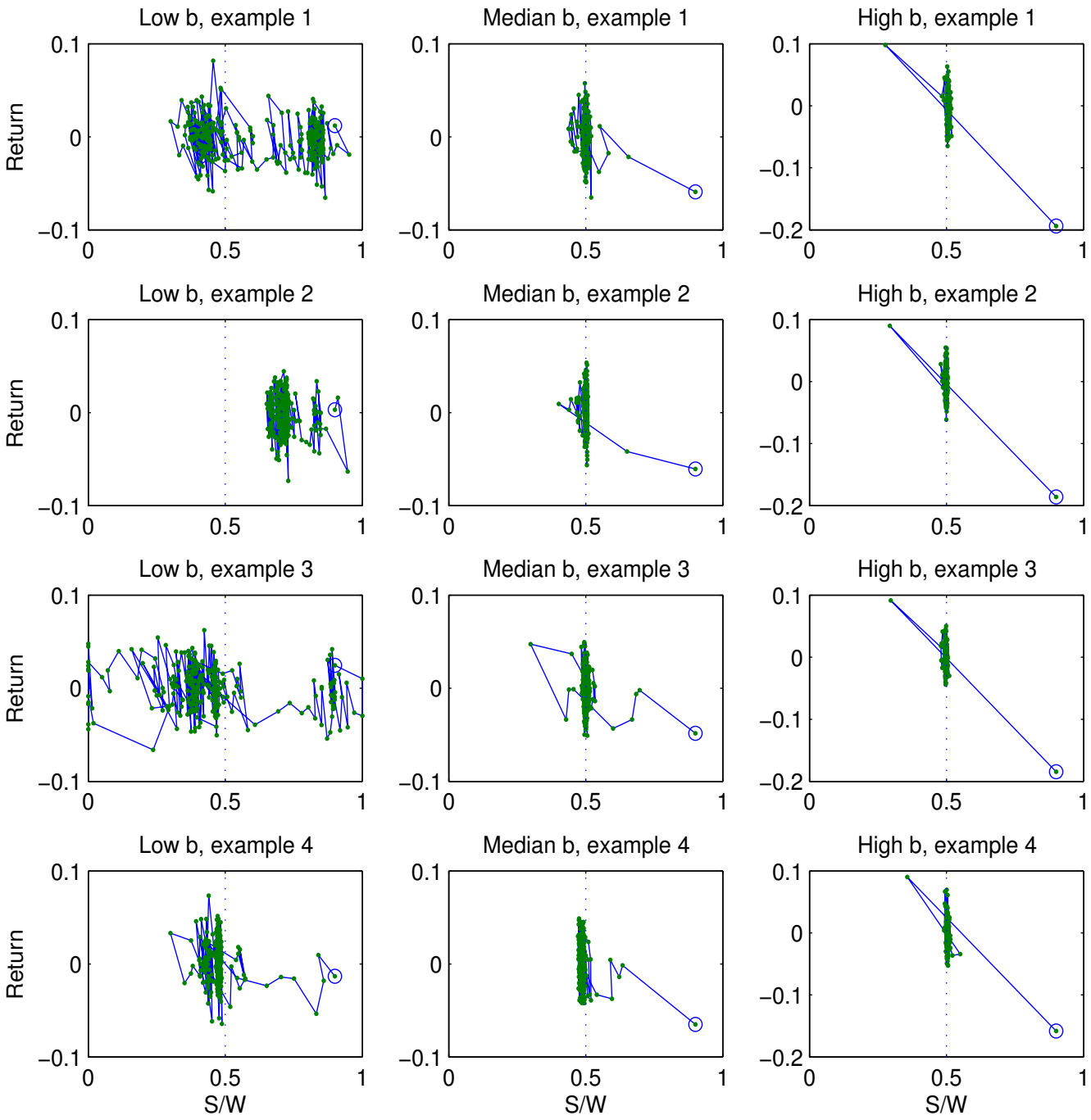

Figure 5. Examples of learning paths. This figure plots representative examples of learning paths for various random samples under the $b \geq 0$ prior. Each of the 12 panels plots aggregate active fund returns $r_{A,t}$ against the aggregate allocation to the active industry $(S/W)_t$ for $t = 1, \ldots, 300$ 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 $S/W = 0.5$. 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 $(S/W)_1 = 0.9$.

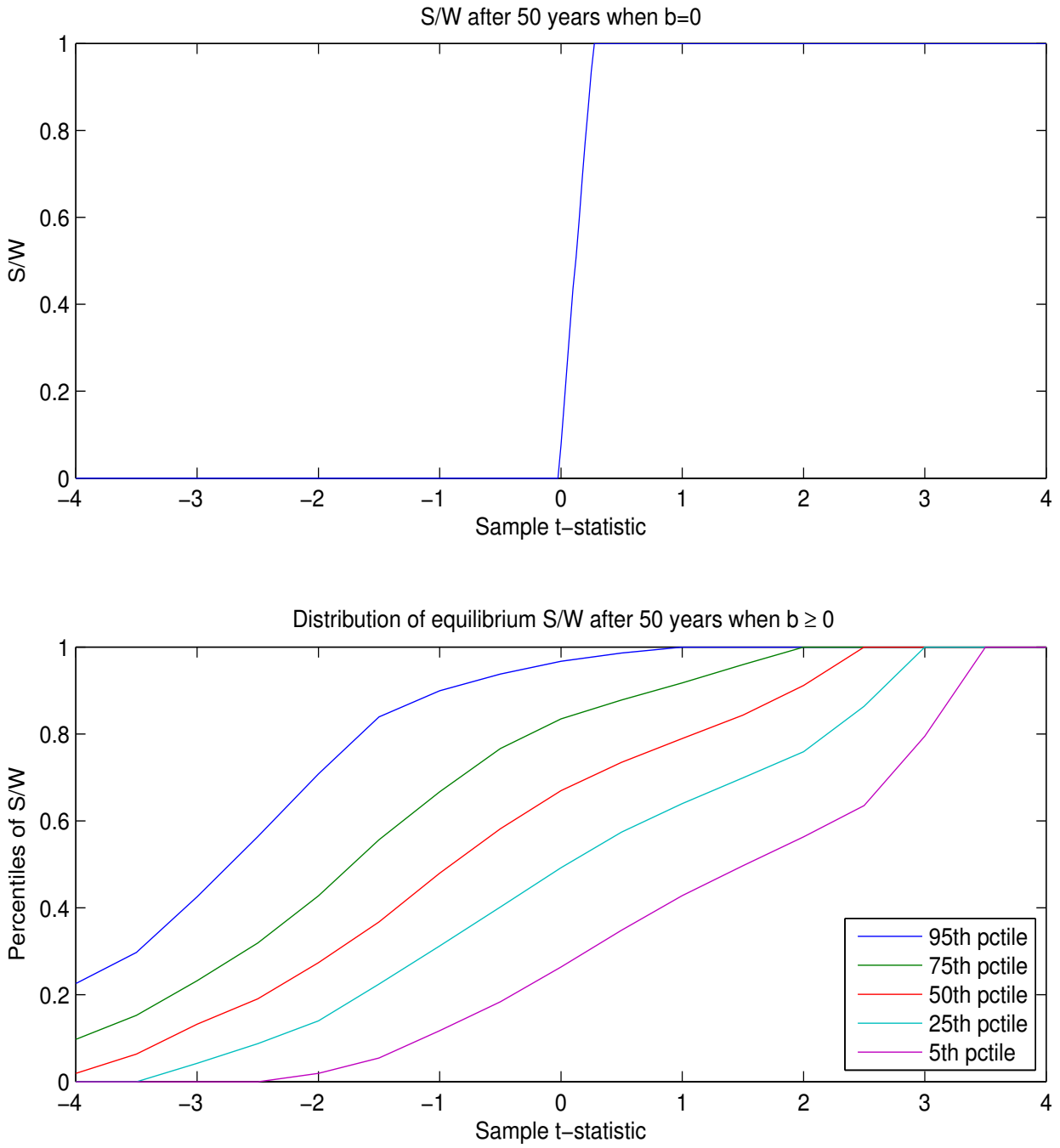

Figure 6. The posterior distribution of the equilibrium allocation to active management conditional on past performance. This figure plots selected percentiles of the posterior distribution of S/W, the equilibrium allocation to active management, conditional on the t-statistic associated with the industry's 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.

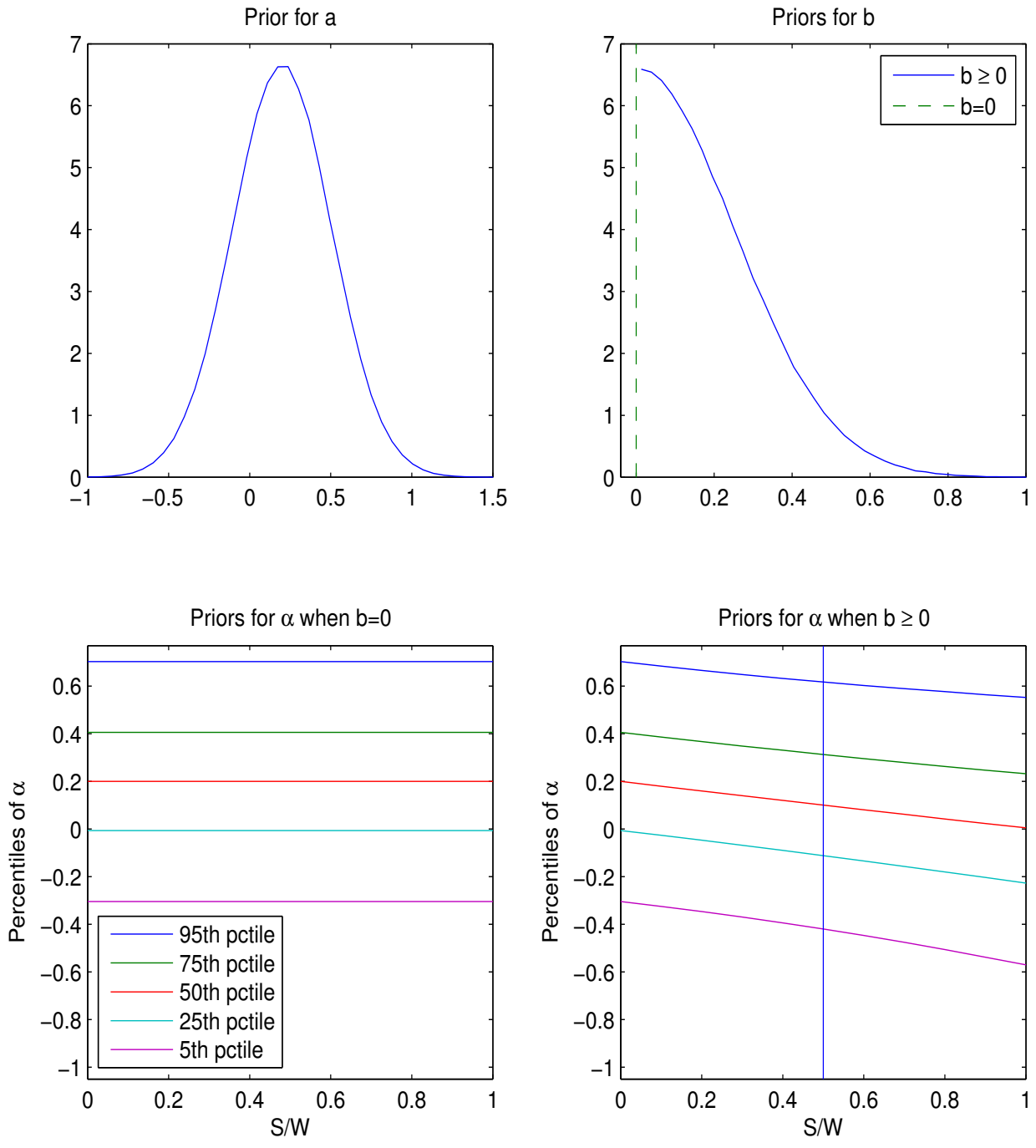

Figure 7. Alternative prior distribution. This figure plots alternative prior distributions for a, b, and α. The priors for b in Panel B are the same as in the baseline prior in Figure 2, but the prior for a in Panel A is more pessimistic. This alternative prior assigns a 26% probability to the event that $a < 0$, which is substantially larger than the 7.2% probability in Figure 2. The prior is chosen such that the initial equilibrium allocation to active management is $S/W = 0.5$ instead of 0.9 as in the baseline prior under decreasing returns to scale. Panels C and D plot the selected percentiles of the implied prior distributions for α as a function of S/W under the priors $b = 0$ and $b \geq 0$, respectively.

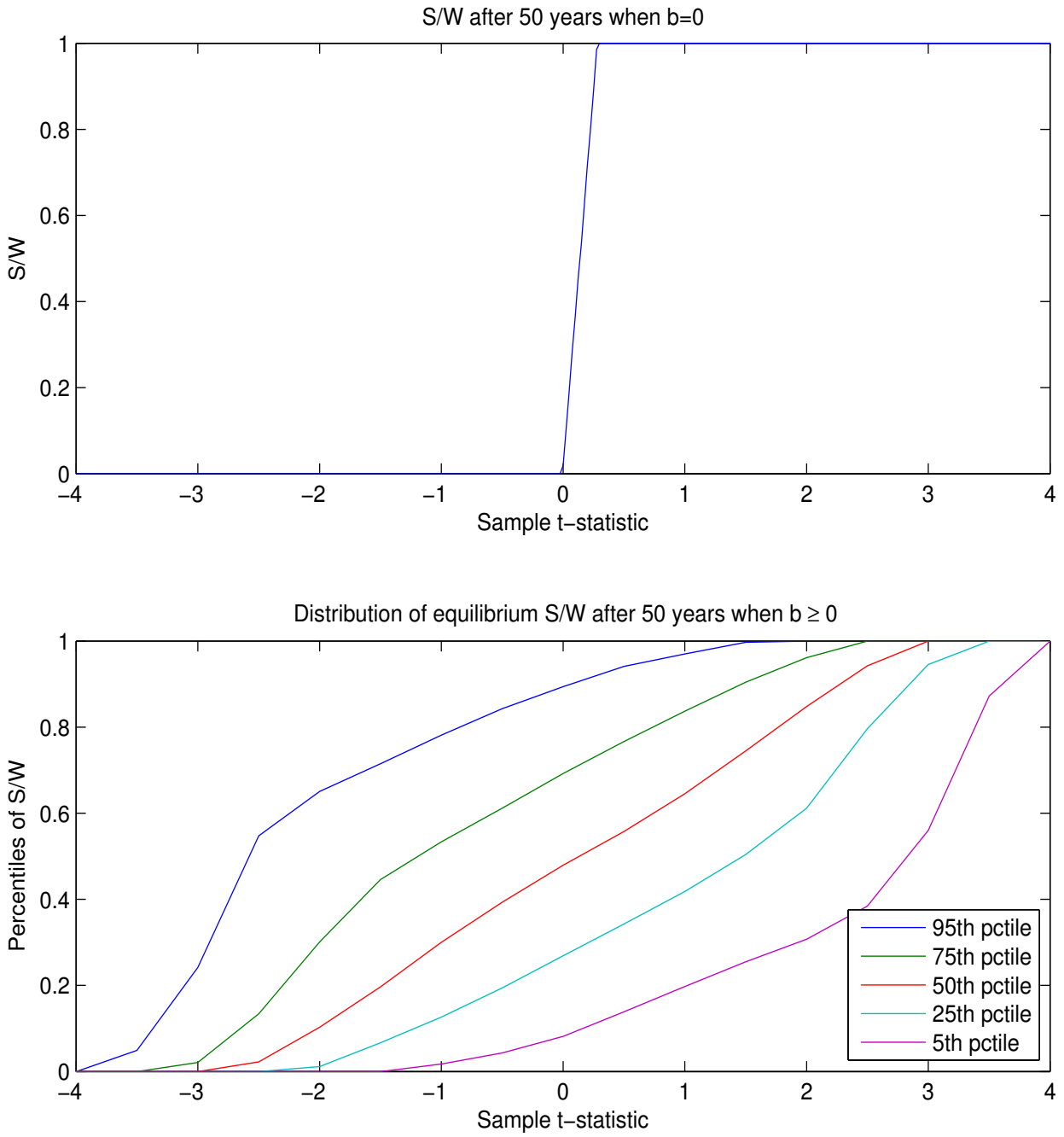

Figure 8. The posterior distribution of the equilibrium allocation to active management conditional on past performance under the alternative prior. This figure plots selected percentiles of the posterior distribution of S/W, the equilibrium allocation to active management, conditional on the t-statistic associated with the industry's historical alpha computed over a period of $T = 50$ years. The figure is analogous to Figure 6, except that the prior from Figure 2 is replaced by the prior from Figure 7. 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.

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.

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.

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.

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.