On the Size of the Active Management Industry

Ľuboš Pástor

University of Chicago and National Bureau of Economic Research

Robert F. Stambaugh

University of Pennsylvania and National Bureau of Economic Research

We argue that active management's popularity is not puzzling despite the industry's poor track record. Our explanation features decreasing returns to scale: As the industry's size increases, every manager's ability to outperform passive benchmarks declines. The poor track record occurred before the growth of indexing modestly reduced the share of active management to its current size. At this size, better performance is expected by investors who believe in decreasing returns to scale. Such beliefs persist because persistence in industry size causes learning about returns to scale to be slow. The industry should shrink only moderately if its underperformance continues.

I. Introduction

Active asset management remains popular, even though its track record has long been unimpressive. For example, consider actively managed eq-

We are grateful for comments from Andrew Ang, Amil Dasgupta, Lord John Eatwell, Gene Fama, Vincent Glode, Will Goetzmann, Rick Green, Rich Kihlstrom, Ralph Koijen, Kim Min, Dimitris Papanikolaou, Monika Piazzesi, Luke Taylor, Rob Vishny, Guofu Zhou, three anonymous referees, workshop participants at Chicago, Drexel, Emory, Michigan State, Ohio State, Temple, and Wharton, as well as participants in the meetings of the National Bureau of Economic Research Asset Pricing Program, Western Finance Association, European Finance Association, Cambridge/Penn conference, HEC Finance and Statistics Conference, Institutional Investor conference at the University of Texas at Austin, CFA Institute Annual Conference, and Q-Group. Support as an Initiative for Global Markets Visiting Fellow (Stambaugh) at the University of Chicago is gratefully acknowledged.

[*Journal of Political Economy*, 2012, vol. 120, no. 4]

uity mutual funds, which constitute a large and well-researched segment of the active management industry. Numerous studies report that these funds have provided investors with average returns significantly below those on passive benchmarks.[1] While this track record could help explain the growth of index funds, the total size of index funds is still modest compared to that of actively managed funds.[2] Given the negative track record, one might be puzzled by the enormous size of the active management industry.

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

$$\alpha_t = a - b\left(\frac{S}{W}\right)_t, \tag{1}$$

where α_t is the industry's expected return at time t in excess of passive benchmarks and $(S/W)_t$ is the industry's size as a fraction of the total amount managed actively and passively. Decreasing returns to scale are captured by $b > 0$. If the benchmarks are sufficient for pricing assets in an efficient market, α_t reflects asset mispricing. In that case, our modeling of decreasing returns to scale is equivalent to assuming that mispricing is reduced as more money seeks to exploit it.

Decreasing returns to scale help us understand the continued popularity of active management. Investors are uncertain about the industry's alpha, and they learn about it from realized returns. After observing negative performance, investors infer that α is lower than expected, and they reduce their allocation to active management. Indeed, the growth of indexing over the past few decades has modestly reduced the share of active management to its current size. The fact that the reduction in S/W has been modest is consistent with the cushioning provided by decreasing returns to scale: a lower S/W implies a higher α going forward. Investors infer that α is too low at the current level of S/W, but they know that α will go up after they reduce S/W, so they disinvest less than they would if returns to scale were constant. Under decreasing returns to scale, past un-

[1] See Jensen (1968), Malkiel (1995), Gruber (1996), Wermers (2000), Pástor and Stambaugh (2002*b*), Fama and French (2010), Del Guercio and Reuter (2011), and others. Fama and French report that, over the past 23 years, an aggregate portfolio of US equity mutual funds underperformed various benchmarks by about 1 percent per year.

[2] The Investment Company Institute (2009) reports that assets of equity mutual funds total $3.8 trillion at the end of 2008. They also report that 87 percent of those assets are under active management, as opposed to being index funds.

derperformance does not imply future underperformance; it implies only that investors should allocate less to active management. After a period of underperformance, the optimal allocation to active management should be smaller than it was at the beginning of the period, but it may remain substantial.

To explore the quantitative implications of the above story, we develop a model of active management featuring Sharpe ratio–maximizing investors and fee-maximizing fund managers. We model decreasing returns to scale in a way similar to equation (1), with unknown parameters a and b. We derive the model's implications for the equilibrium size of the active management industry, measured in relative terms as S/W. We also solve for the equilibrium α and the manager fee.

We find that the industry's equilibrium size depends critically on the degree of competition among investors and fund managers. The role of competition is especially clear in the special case in which investors are risk neutral. In the absence of competition among either investors or managers, the equilibrium industry size maximizes the expected total profit. If investors compete but managers do not, all the profit goes to managers in the form of fees; if managers compete but investors do not, all the profit goes to investors in the form of alpha. A different picture emerges under perfect competition among both investors and managers. Interestingly, the industry's fully competitive equilibrium size is twice as large as the size obtained if either type of competition is shut down. The fully competitive industry produces zero expected total profit, so that investors earn zero alpha and managers earn no abnormal fee.

Our results highlight an externality that is inherent in active investing under decreasing returns to scale: when investors compete, they dilute each other's returns by investing to the point at which the expected active alpha is zero. Owing to this externality, competition results in overproduction of active management relative to the profit-maximizing size, making it easier to understand why active management is so popular. If more active management implies less mispricing, then more competition also implies more efficient asset markets. This result has clear policy implications.

Focusing on the fully competitive setting, we compare the model-implied equilibrium size of the active management industry with the actual size. To measure the industry's actual size, we rely on data for US equity mutual funds, which we assume to be representative of the industry as a whole. The advantage of using mutual fund data is that the histories of fund returns and assets under management are longer and more reliable than those of any other segment of the asset management industry. We measure the industry's actual S/W as assets under management for all active funds divided by assets under management for active and passive funds combined. The latest value of this ratio, computed as of the beginning of 2006, is 0.87.

We examine the conditions under which the rational investors in our model currently choose an 87 percent allocation to active management. The investors choose their allocations after updating their prior beliefs about a and b with the available historical data. The data consist of a 44-year history of the actual S/W values and returns on the aggregate portfolio of actively managed US equity mutual funds. This history paints an unfavorable view of active funds, whose aggregate portfolio has significantly underperformed the market.[3] Despite the negative return history, we find that the 87 percent allocation to active management is consistent with a variety of prior beliefs, as long as those beliefs feature decreasing returns to scale. For example, for our baseline prior specification for a, the 87 percent allocation is chosen by investors who expect b to be about 0.1 a priori. Our results seem robust to alternative prior specifications. We conclude that the observed large size of the active management industry can be rationalized by decreasing returns to scale in the industry.

In contrast, active management's popularity would seem quite puzzling under the more traditional assumption of constant returns to scale ($b = 0$ in eq. [1]). This assumption is routinely adopted by performance evaluation studies, in which alphas are generally treated as constants, unrelated to the industry's size. We find that under constant returns to scale, the current size of the active management industry should be zero. With $b = 0$, the industry's track record quickly leads investors to perceive $\alpha < 0$ at any S/W, even if their prior beliefs about α are more optimistic than those leading to the results mentioned above under decreasing returns to scale. With $\alpha < 0$, any positive investment in active management would be undesirable for mean-variance investors; they would instead go short if they could. If our rational investors thought returns to scale were constant, the active management industry would have disappeared many years ago.

Is the industry likely to remain large in the future? To answer this question, we simulate future paths of returns from our model under prior beliefs that are consistent with the industry's current size. We then calculate the expected future industry size after observing various potential track records. We find that S/W is likely to remain large for a long time, even if the industry continues to significantly underperform its benchmark. For example, conditional on the future t-statistic of alpha equal to -2, S/W is expected to decline only to 63.4 percent after 20 years. The industry's decline in response to underperformance is restrained by decreasing returns to scale: investors know that when they allocate less to active management, their future active returns will be higher. In contrast, the industry would shrink much faster in response to underperfor-

[3] When we regress aggregate active fund excess returns on market excess returns in our full sample of January 1962 through September 2006, the annualized estimated alpha is −88 basis points, with a t-statistic of −2.7. Our data, which come from Ken French, are described in more detail in Sec. III.B.

mance if returns to scale were constant: for example, for the same t-statistic of -2, the industry would disappear after just 1 year of underperformance, which seems implausible. We conclude that owing to decreasing returns to scale, the active management industry is likely to remain large for many years.

Our proposed reconciliation of the active management industry's large size with its poor track record is the main contribution of this paper. Our second contribution is to show that learning about returns to scale in active management is slow. Investors in our model face endogeneity that limits their learning about a and b in equation (1). As investors update their beliefs about a and b, they adjust S/W. They learn about a and b by observing the industry's returns that follow different allocations. The extent to which they learn is thus endogenous: what they learn affects how much they allocate, but what they allocate affects how much they learn. At the extreme, if investors were to keep S/W constant over time, they would eventually learn the value of α at that level of S/W, but they would learn nothing about a and b individually. While the equilibrium S/W generally does vary over time, its fluctuations are significantly muted by decreasing returns to scale. This lack of variation in S/W impedes learning about a and b. As a result, investors remain highly uncertain about a and b even after observing long histories of data. For the same reason, the investors' initial beliefs about returns to scale persist for a long time.

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, whereas 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 "active-management puzzle." Departing from Berk and Green's cross-sectional focus, we analyze the aggregate size of the active management industry.

We are not alone in trying to explain the puzzling popularity of active management in light of its poor track record. 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 (2011) presents an explanation in which investors expect negative future performance as a fair trade-off 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 (2002*a*) 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 relates to a number of other directions in recent research. Viewed broadly, the study adds to a growing literature addressing the size of various aspects of the financial industry (e.g., Philippon 2008; Bolton, Santos, and Scheinkman 2011). Garcia and Vanden (2009) 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 that in Berk and Green (2004).[4] Recent models of mutual fund formation also include Mamaysky and Spiegel (2002) and Stein (2005). Neither these models nor that of Garcia and Vanden examines 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, Khorana, Servaes, and Tufano (2005) empirically analyze the determinants of the size of the mutual fund industry across countries.

The paper is organized as follows. Section II presents our model. After describing the general setting, we first examine the case in which investors are risk neutral. The simple results obtained there for alphas, fees, and industry size clearly reveal the role of competition among managers and investors. We then move to a mean-variance setting, which forms the basis

[4] 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 Vayanos and Woolley (2008), Petajisto (2009), Cuoco and Kaniel (2011), Dasgupta, Prat, and Verardo (2011), Guerrieri and Kondor (2012), and He and Krishnamurthy (forthcoming).

for our empirical work. Section III discusses the priors and their updating with data. Section IV presents the model's quantitative implications for the industry's current size given its historical track record. Section V calculates the expected future industry size after observing various potential future track records. It also discusses the properties of learning about returns to scale. Section VI relates our model to that of Berk and Green (2004). Section VII presents conclusions.

II. Model

A. *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. We focus primarily on a perfectly competitive setting with infinite numbers of both managers and investors who play infinitesimal individual roles in equilibrium ($M \to \infty$, $N \to \infty$). To emphasize the important role of competition when there are decreasing returns to scale, we also consider two alternative settings. In one there is perfect competition among managers but only a single investor ($M \to \infty$, $N = 1$), whereas in the other there is perfect competition among investors but only a single manager ($M = 1$, $N \to \infty$).

The rates of return earned by investors in the managers' funds obey the regression model

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

where r_F is the $M \times 1$ vector of fund returns in excess of the riskless rate, $\underline{\alpha}$ is the $M \times 1$ vector of fund alphas, r_P is the excess return on the passive benchmark portfolio, β is the $M \times 1$ vector of fund betas, and u is the $M \times 1$ vector of the residuals. The passive benchmark portfolio's excess return has mean μ_P and variance σ_P^2. We suppress time subscripts throughout to simplify notation. 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, a variance of σ_ϵ^2, and zero correlation with each other. The common factor x has mean zero and variance σ_x^2. The values of β, μ_P, σ_P, 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 equal to S plus the amount invested in the passive benchmark. 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. 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 parameters a and b in equation (4) are unknown. We denote their first and second conditional moments by

$$\mathrm{E}\left(\begin{bmatrix} a \\ b \end{bmatrix} \middle| D\right) = \begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix}, \tag{5}$$

$$\mathrm{Var}\left(\begin{bmatrix} a \\ b \end{bmatrix} \middle| D\right) = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}, \tag{6}$$

where D denotes the set of information available to investors.

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

The parameter b determines the degree to which the expected benchmark-adjusted return for any manager declines as the relative size of active management increases. We allow $b \geq 0$, although it is likely that $b > 0$ because of decreasing returns to scale in the active management industry.

[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.

As more money chases opportunities to outperform, prices are affected, and such opportunities become more difficult for any manager to identify. Prices are affected by these profit-chasing actions of active managers unless markets are perfectly liquid. In that sense, b is related to market liquidity: $b = 0$ in infinitely liquid markets but $b > 0$ otherwise.

We specify the relation (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,

$$\max_{f_i} f_i s_i. \tag{7}$$

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

$$\alpha_i = a - b\frac{S}{W} - f_i. \tag{8}$$

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

The N investors are assumed to allocate between the active funds and the benchmark portfolio so as to maximize the Sharpe ratio of the resulting combination. Let δ_j denote the $M \times 1$ vector of the weights that investor j places on the M funds. For each investor j the allocations to the funds solve the problem

$$\max_{\delta_j}\left\{\frac{\mathrm{E}(r_j|D)}{\sqrt{\mathrm{Var}(r_j|D)}}\right\}, \tag{9}$$

where the excess return on the investor's portfolio is given by

$$r_j = \delta_j' r_F + \left(1 - \delta_j' \iota_M\right) r_P, \tag{10}$$

and ι_M denotes an M-vector of ones. We impose the restriction that all elements of the $M \times 1$ vector δ_j are nonnegative (no shorting of funds).

In equilibrium, the proportional allocation to active management chosen by each investor is equal to the aggregate ratio S/W. The equilibrium

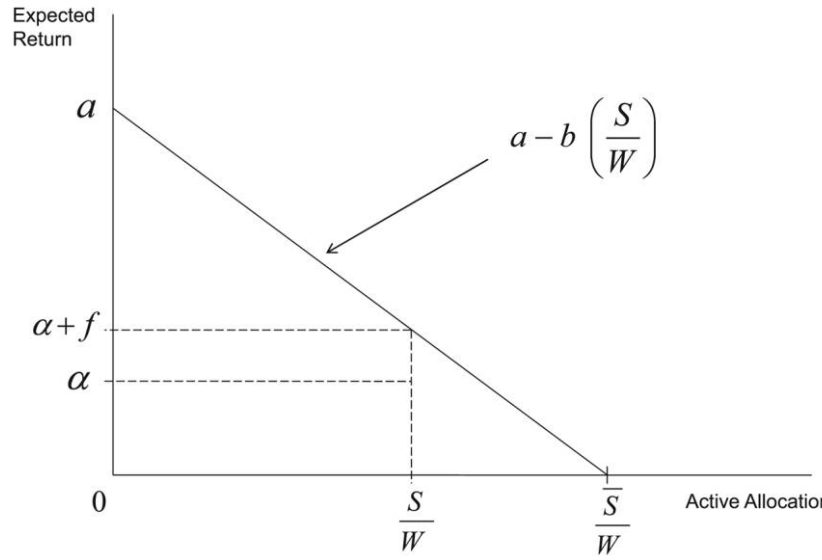

FIG. 1.—Decreasing returns to scale for the active management industry. This figure plots the theoretical relation between the expected benchmark-adjusted excess fund return before fees against the relative size of the active management industry. Specifically, it plots equation (8): $\alpha + f = a - b(S/W)$, where α is the expected benchmark-adjusted excess fund return earned by investors, f is the proportional fee charged by the fund manager, and S/W is the aggregate allocation to active management. For $b > 0$, the industry exhibits decreasing returns to scale. The values of α, f, and S/W are determined in equilibrium. At $S/W = \bar{S}/W$, we have $\alpha = f = 0$.

is partial in several respects. The benchmark portfolio's returns are assumed to be exogenously given and, thus, unaffected by the actions of investors and fund managers. In addition, the managers' potential outperformance comes at the expense of other investors whose decisions are not modeled here.[6] We also isolate an investor's active-versus-passive allocation decision from his labor income, real estate, nationality, and any state variables that are typically associated with hedging demands. While such variables may well be relevant for the investor's consumption/investment decision and

[6] The latter investors are required by the fact that alphas (before costs) must aggregate to zero across all investors (see, e.g., Sharpe 1991; Fama and French 2010). In the absence of such other investors, one could not expect active managers to earn positive alphas. These other investors might trade for exogenous "liquidity" reasons, e.g., or they could engage in their own active (nonbenchmark) 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 et al. 2006). Good candidates for such investors are individuals who invest in financial markets directly. For example, the proportion of US equity held directly by individuals is substantial: in 1980–2007, this proportion ranged from 22 percent in 2007 to 48 percent in 1980 (French 2008).

even his overall asset allocation decision, they seem unlikely to have first-order effects on deciding how to split financial wealth between active and passive funds. In particular, it is not clear why any state variables should be related to the nonbenchmark risk in active management, represented by u in (2).

We assume that all funds have a beta of one: $\beta = \iota_M$, where β is defined in equation (2).[7] Combined with equations (2) and (10), this assumption gives

$$r_j = r_P + \delta_j'(\underline{\alpha} + u). \tag{11}$$

The benchmark return r_P is then present in the investor's portfolio return for any choice of δ_j. In other words, with this unit-beta assumption, the allocation decision hinges on the funds' active contributions, $\underline{\alpha} + u$, but not on their benchmark exposures.

B. *Equilibrium under Risk Neutrality*

Before turning to the mean-variance objective in (9), we first analyze the risk-neutral setting in which investors simply maximize expected return, solving the problem

$$\max_{\delta_j} \{\mathrm{E}(r_j|D)\}. \tag{12}$$

This simpler setting allows a more transparent analysis of the effects of competition that also arise in the mean-variance setting, as shown later in Section IV. The following proposition gives the equilibrium values of the key quantities under three alternative specifications of the nature of competition between managers and investors.

PROPOSITION 1. In equilibrium for investors and managers when $\tilde{a} > 0$, we have $\mathrm{E}(\alpha_i|D) = \tilde{\alpha}$ and $f_i = f$ for all funds i receiving positive investment. With perfect competition among both managers and investors ($M \to \infty$, $N \to \infty$),

$$f = 0, \tag{13}$$

$$\tilde{\alpha} = 0, \tag{14}$$

$$\frac{S}{W} = \frac{\tilde{a}}{\tilde{b}}. \tag{15}$$

[7] This assumption is consistent with empirical evidence for active equity mutual funds. On the basis of monthly data for the January 1962–September 2006 period, the aggregate portfolio of US actively managed mutual funds has a beta of 0.99 with respect to the value-weighted market index.

With perfect competition among managers but only a single investor ($M \to \infty$, $N = 1$),

$$f = 0, \tag{16}$$

$$\tilde{\alpha} = \frac{\tilde{a}}{2}, \tag{17}$$

$$\frac{S}{W} = \frac{\tilde{a}}{2\tilde{b}}. \tag{18}$$

With perfect competition among investors but only a single manager ($M = 1$, $N \to \infty$),

$$f = \frac{\tilde{a}}{2}, \tag{19}$$

$$\tilde{\alpha} = 0, \tag{20}$$

$$\frac{S}{W} = \frac{\tilde{a}}{2\tilde{b}}. \tag{21}$$

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

Proof. See the Appendix.

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

With competing investors, the equilibrium expected alpha is $\tilde{\alpha} = 0$. Each investor in that setting sees his own investment as having no effect on S/W and thus no effect on alphas. Investors impose a negative externality on each other: they dilute each other's returns by investing to the point at which the expected alpha on all active funds is zero. If the ex-

pected alpha were instead positive, all investors would be dissatisfied with their current holding of active funds and would wish to increase it, thereby raising S/W and lowering alpha.

With no competition among managers or no competition among investors, the industry size is only half as large as in the fully competitive setting (compare eq. [15] with [18] and [21]). The value in (18) and (21) is also the value that maximizes expected total profit. That is, from equation (4), expected total profit is

$$\Pi = \sum_{i=1}^{M} \pi_i = S\left(\tilde{a} - \tilde{b}\frac{S}{W}\right), \tag{22}$$

which is maximized at $S/W = \tilde{a}/(2\tilde{b})$, the value in (18) and (21). At that value, $\Pi = S(\tilde{a}/2)$, equivalent to an expected rate of return of $\tilde{a}/2$ on the invested amount S. That expected rate of return, $\tilde{a}/2$, is earned as $\tilde{\alpha}$ in (17) by a single investor when facing competing managers. The same rate of $\tilde{a}/2$ is also charged as the fee in (19) by a single manager facing competing investors. With $N = 1$, profit is maximized because the single investor fully internalizes the fact that his own investment determines S/W and, thereby, expected profit. With $M = 1$, the single manager acts as a monopolist in setting the fee such that the resulting industry size produces fee revenue that captures the maximum expected profit.

Competition makes it easier to understand why the active management industry is large. When both managers and investors compete, the resulting industry size of $S/W = \tilde{a}/\tilde{b}$ in (15) is twice as large as the profit-maximizing size, as noted earlier. This fully competitive industry size, which is denoted by $\overline{S}/W$ in figure 1, produces zero expected profit in (22). Despite generating no profits for investors or managers, the fully competitive industry can nevertheless provide a positive externality to asset markets. Suppose that the benchmark is "correct" in an asset-pricing context, in that securities with nonzero alphas with respect to the benchmark are mispriced. Opportunities to outperform the benchmark then reflect mispricing. If no money actively chased mispricing ($S = 0$), some mispricing would likely exist. By moving prices toward fair values, the industry provides a positive externality to the (unmodeled) real side of the economy.

In the maximization in (12), we impose the lower bound of zero on the elements of δ_j, but we have not imposed an upper bound. A reasonable alternative is to impose the constraint

$$\delta_j' \iota_M \leq 1, \tag{23}$$

which precludes shorting of the passive benchmark portfolio (cf. [10]). When (23) binds, S/W in equation (15), (18), or (21) exceeds one, and

the constrained equilibrium value of S/W instead equals one. Also, as in the earlier unconstrained setting, $f = 0$ with competition among managers, but then $\tilde{\alpha} = \tilde{a} - \tilde{b}$, a positive value that does not depend on whether investors compete. In essence, the constraint in (23) then prevents investors from increasing the size of the industry to the point at which all profit is eliminated. In contrast, when there is just a single manager and (23) binds, the manager earns a fee greater than the value in (19), while competition among investors still delivers $\tilde{\alpha} = 0$. The Appendix includes a treatment of the case in which (23) binds.

C. *Equilibrium in the Mean-Variance Setting*

We now turn to the mean-variance setting in which investors maximize the objective function in (9). Our primary focus is on the fully competitive case ($M \to \infty$, $N \to \infty$). In Section IV, we also discuss results under the additional scenarios discussed above, in which either $M = 1$ or $N = 1$. Those results show that the equilibrium values of S/W closely follow the same relative proportions across the alternative scenarios as in the risk-neutral setting. That is, with either a single investor ($N = 1$, $M \to \infty$) or a single manager ($M = 1$, $N \to \infty$), S/W is only about half as large as its fully competitive value. Unlike the fully competitive case, where the equilibrium can be computed analytically, the two additional cases require numerical solutions. The solution procedures for those cases are explained in the Appendix.

The explicit analytic solution for S/W in the fully competitive case—the solution to a cubic equation—is fairly cumbersome. We instead simply present that cubic equation in the following proposition.

PROPOSITION 2. In equilibrium with perfect competition among both managers and investors, if $\tilde{a} > 0$, then S/W is given by the (unique) real positive solution to the equation

$$0 = \tilde{a} - \frac{S}{W}\left[\tilde{b} + \gamma\left(\sigma_a^2 + \sigma_x^2\right)\right] + \left(\frac{S}{W}\right)^2 2\gamma\sigma_{ab} - \left(\frac{S}{W}\right)^3 \gamma\sigma_b^2 \qquad (24)$$

when the constraint in (23) does not bind, where $\gamma = \mu_P/\sigma_P^2$. If investors also face the constraint in (23) and the solution to (24) exceeds one, then $S/W = 1$. If $\tilde{a} \leq 0$, then $S/W = 0$.

Proof. See the Appendix.

When the equilibrium value of S/W lies between zero and one, it can be represented in mean-variance terms. To see this, let r_A denote the benchmark-adjusted return on the aggregate portfolio of all funds:

$$\begin{aligned} r_A &= \frac{1}{M}\iota_M' r_F - r_P \\ &= \frac{1}{M}\iota_M' \underline{\alpha} + x + \frac{1}{M}\sum_{i=1}^{M} \epsilon_i \\ &= a - b\frac{S}{W} + x + \frac{1}{M}\sum_{i=1}^{M} \epsilon_i, \end{aligned} \tag{25}$$

using equations (2), (3), (8), and the result that $f = 0$ in equilibrium. Thus, as $M \to \infty$,

$$r_A = a - b\frac{S}{W} + x \tag{26}$$

since the variance of the last term in (25) goes to zero. It follows from (26) that

$$\mathrm{E}(r_A|D) = \tilde{a} - \tilde{b}\frac{S}{W} \tag{27}$$

and

$$\mathrm{Var}(r_A|D) = \sigma_a^2 + \sigma_x^2 - 2\left(\frac{S}{W}\right)\sigma_{ab} + \left(\frac{S}{W}\right)^2\sigma_b^2. \tag{28}$$

Equation (24) can then be rewritten as

$$\frac{S}{W} = \frac{\tilde{a} - \tilde{b}(S/W)}{\gamma\left[\sigma_a^2 + \sigma_x^2 - 2(S/W)\sigma_{ab} + (S/W)^2\sigma_b^2\right]} \tag{29}$$

$$= \frac{\mathrm{E}(r_A|D)}{\gamma \mathrm{Var}(r_A|D)}, \tag{30}$$

where the resulting mean-variance expression in (30) relies on (27) and (28).

We can also write equation (26) as $r_A = \alpha + x$, with $\alpha = a - b(S/W)$, so that $\mathrm{Var}(r_A|D) = \sigma_x^2 + \sigma_\alpha^2$, where $\sigma_\alpha^2 = \mathrm{Var}(\alpha|D)$. Equation (30) can then be rewritten as

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

which gives

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

Note that σ_α^2 depends on S/W, thus requiring the solution to the cubic equation in (24). In the special case in which a and b are known, $\sigma_\alpha^2 = 0$ and the right-hand side of (32) yields the solution directly, so solving the cubic equation is then unnecessary. As before in the risk-neutral solution (15), we see in (32) that greater profitability of the first dollar invested (higher $\tilde{a}$) makes the equilibrium industry size larger while more strongly decreasing returns to scale (higher $\tilde{b}$) makes the industry smaller.

The role of uncertainty about a and b in determining industry size can be seen in (32). This uncertainty enters through uncertainty about α, which enters the denominator in (32) via $\gamma\sigma_\alpha^2$. Greater uncertainty about α thus makes the industry smaller. We specify γ $(= \mu_P/\sigma_P^2)$ as 1.92, which is based on estimates for the market portfolio and the same January 1962–September 2006 period over which our fund data are available.[8] With this value of γ, the product $\gamma\sigma_\alpha^2$ can exert a nontrivial effect on industry size when there is substantial uncertainty about a and b, such as one might possess before updating prior beliefs with data. After such updating, however, the magnitude of $\gamma\sigma_\alpha^2$ is often small compared to $\tilde{b}$, which also appears in the denominator in (32). Thus, posterior uncertainty about a and b typically does not exert a large effect on the equilibrium S/W.

Uncertainty about the unexpected active return also affects industry size, via $\gamma\sigma_x^2$ in the denominator in (32). We specify the volatility of the aggregate active benchmark-adjusted return as $\sigma_x = 0.02$, or 2 percent per year, which is approximately equal to the annualized residual standard deviation from the regression of the value-weighted average return of all active US equity mutual funds on the market benchmark in the January 1962–September 2006 period. With this value of σ_x, the value of $\gamma\sigma_x^2$ is often small compared to $\tilde{b}$, as is the value of $\gamma\sigma_\alpha^2$ after prior beliefs are updated with data. As a result, after updating with data, the equilibrium value of S/W is generally well approximated by (15), which omits the term $\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)$ that appears in (32).

The industry's expected alpha, $\tilde{\alpha} = \mathrm{E}(r_A|D)$, can be obtained by combining equations (27) and (32) to give

$$\tilde{\alpha} = \tilde{a}\left[\frac{\gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)}{\tilde{b} + \gamma\left(\sigma_x^2 + \sigma_\alpha^2\right)}\right]. \tag{33}$$

We see from (33) that $\tilde{\alpha} > 0$. In order for a positive allocation to active management to offer investors a higher Sharpe ratio than the passive benchmark, investors must expect compensation for the nondiversifiable risky component x as well as for uncertainty about alpha.

[8] Our data are described in more detail in Sec. III.B.

III. Prior and Posterior Beliefs

In this section, we discuss the prior and posterior beliefs about the key parameters, a and b, as well as beliefs about the alpha implied by those parameters. We also describe our data.

A. *Prior Beliefs*

A common assumption in the literature is that returns to scale in active investing are constant ($b = 0$). While we consider such a dogmatic prior belief as well, our focus is on prior beliefs in which returns are decreasing in scale at an uncertain rate (i.e., b is an unknown positive value). We show in Section IV that investors who believe a priori that $b > 0$ make very different investment decisions than investors who believe that $b = 0$, even after observing exactly the same evidence.

To capture decreasing returns to scale, we specify a bivariate normal joint prior distribution for a and b, truncated to require that $b \geq 0$:

$$\begin{bmatrix} a \\ b \end{bmatrix} \sim N(E_0, V_0) I(b \geq 0), \tag{34}$$

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 one if condition c is true and zero otherwise. Denote

$$E_0 = \begin{bmatrix} E_0^a \\ E_0^b \end{bmatrix}, \quad V_0 = \begin{bmatrix} V_0^{aa} & V_0^{ab} \\ V_0^{ab} & V_0^{bb} \end{bmatrix}. \tag{35}$$

We specify $E_0^b = V_0^{ab} = 0$ for simplicity. We consider a wide range of prior means of b, denoted by b_0. In this section, we focus on $b_0 = 0.1$, a value of particular interest in the subsequent analysis. Given the properties of the truncated normal distribution, this prior mean implies $V_0^{bb} = 0.016$ and a prior standard deviation for b equal to $\sigma_b^0 = 0.076$. The marginal prior distribution for b is plotted in panel *B* of figure 2: it is the right half of a zero-mean normal distribution truncated below at zero.

Panel *A* of figure 2 plots three different marginal prior distributions for a. All three distributions are normal. Their means and standard deviations, a_0 and σ_a^0, are specified such that investors with those prior beliefs would optimally choose a given fraction $(S/W)_0$ before observing the historical data. We consider three values of $(S/W)_0$: 0.6, 0.8, and 1.0. For a given value of $(S/W)_0$, there generally exist multiple pairs of (a_0, σ_a^0) for which $(S/W)_0$ is the optimal allocation. To pick a single pair, we impose the additional constraint that the prior probability of $a < 0$ is 1 percent. This constraint is motivated by the discussion presented earlier in Section II. Recall that a represents the expected return on the initial frac-

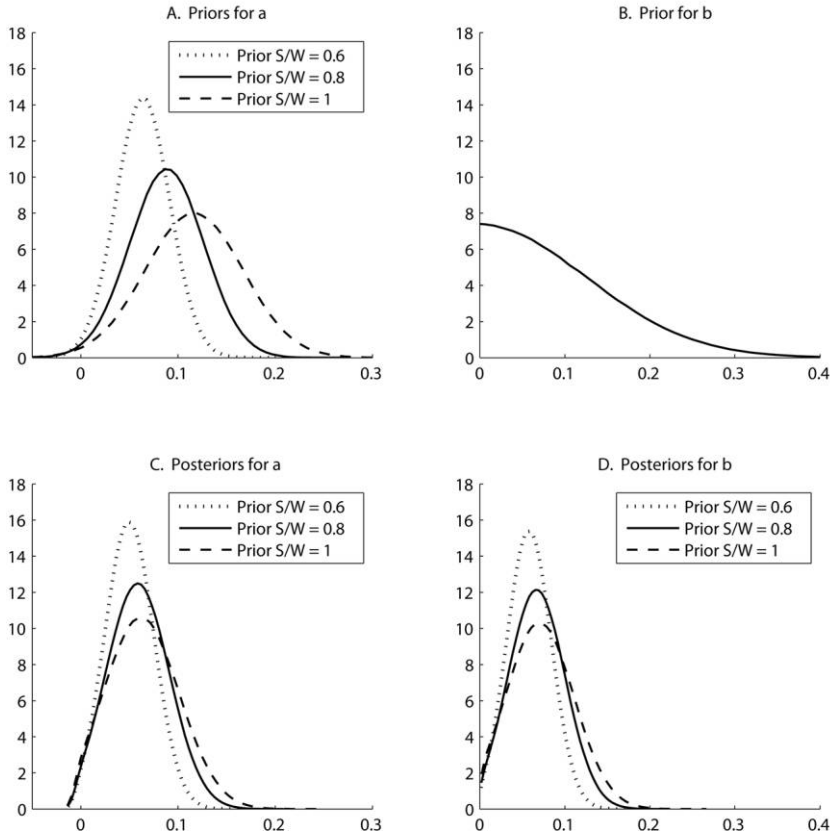

FIG. 2.—Prior and posterior distributions for a and b. The figure plots priors and posteriors for a and b in the function $\alpha_t = a - b(S/W)_t - f$, where α_t is the expected active benchmark-adjusted return, $(S/W)_t$ is the aggregate allocation to active management, and f is the fee (zero under competition). The prior for b shown in panel B has a mean of 0.1. Three priors for a based on that value are shown in panel A. All three have the property prob$(a < 0) = .01$ but differ with respect to the "prior S/W": the S/W under the fully competitive equilibrium based only on prior beliefs. Panels C and D display the posteriors obtained after updating with the mutual fund data, covering the 1962–2005 period.

tion of wealth invested in active management. As we argued in Section II, if no money were invested in active management, some opportunities to outperform passive benchmarks would likely exist, so the initial active investment would almost certainly have a positive expected benchmark-adjusted return. We thus specify priors that admit only a small (1 percent) probability of $a < 0$, as shown in panel A. In the robustness analysis presented in Section IV.D, we also consider probabilities of 0.1 percent and 10 percent.

B. Data

Investors update their prior beliefs with the histories of active returns $\{r_{A,t}\}$ and equilibrium active allocations $\{(S/W)_t\}$. In theory, these quantities can apply to the active management industry in its largest sense, encompassing not only mutual funds but also other segments of the industry, such as defined-benefit pension funds. If the data were available, our empirical analysis could take that broadest perspective. In reality, however, the availability of long series of reliable historical data limits our analysis to mutual funds, a major segment of the industry. We view mutual fund data as a reasonable representation of active management as a whole, in terms of both its historical returns and its share relative to passive investing. It is difficult to observe evidence for or against such a view, but we suggest that it seems reasonable. An alternative interpretation of our model, narrower but also reasonable, is simply that it pertains to mutual funds.

For the series of both $r_{A,t}$ and $(S/W)_t$, we use the data compiled by Fama and French (2010).[9] For each year t from 1963 through 2006, we set $r_{A,t-1}$ equal to the return on the aggregate portfolio of actively managed US equity mutual funds, net of the return attributable to the portfolio's estimated exposures to the Center for Research in Security Prices value-weighted market portfolio. For the market portfolio to represent a fair benchmark for active funds, we need to take into account the small but nontrivial cost of holding the market. We do so by subtracting 15 basis points from each annual market return.[10] For each year t, we also construct $(S/W)_t$ as the ratio of total assets under management for nonindex funds to total assets under management for all funds, both measured at the beginning of year t. As shown in table 1, $(S/W)_t$ equals 1.00 from 1962 through 1976, and then it gradually declines to 0.87 at the end of 2005. The timing is such that $r_{A,t}$ is the return following investors' equilibrium allocation $(S/W)_t$.

C. Posterior Beliefs

To update their beliefs about a and b, investors conduct inference about the coefficients in a time-series regression of returns on the equilibrium allocations. At the end of the sample of T years, the available data

[9] We are grateful to Ken French for providing the data, which end in September 2006. The classification of funds as active vs. passive is performed by Fama and French.

[10] This amount is slightly smaller than the expense ratios of Vanguard's 500 Index Fund and Total Stock Market Index Fund (17 and 18 basis points, respectively, as of 2011), which are among the largest and cheapest index funds available to retail investors. Expense ratios of index funds used to be higher in the funds' early years. Our assumption of the same low expense ratio throughout the sample is conservative in that using a higher index fund expense ratio would make it easier to rationalize a high allocation to active management.

TABLE 1
RETURNS AND RELATIVE SIZE OF ACTIVELY MANAGED FUNDS

Year	S/W	Return	Adjusted Return	Year	S/W	Return	Adjusted Return
1962	1.0000	−16.84	−4.09	1984	.9962	−11.63	−5.03
1963	1.0000	16.05	−.97	1985	.9951	20.43	−2.32
1964	1.0000	11.01	−1.19	1986	.9946	9.60	.66
1965	1.0000	14.68	4.71	1987	.9920	−3.19	.47
1966	1.0000	−9.89	3.24	1988	.9910	9.30	−1.37
1967	1.0000	23.65	.26	1989	.9865	17.64	−1.60
1968	1.0000	5.60	−2.90	1990	.9809	−13.91	−.32
1969	1.0000	−18.99	−2.03	1991	.9769	28.65	1.71
1970	1.0000	−12.56	−6.20	1992	.9718	4.53	−.70
1971	1.0000	13.71	2.44	1993	.9615	11.29	3.04
1972	1.0000	11.08	−1.81	1994	.9579	−5.23	−.58
1973	1.0000	−27.49	−2.54	1995	.9535	26.82	−2.08
1974	1.0000	−34.12	.76	1996	.9391	12.68	−2.60
1975	1.0000	26.89	−3.46	1997	.9251	19.58	−4.51
1976	1.0000	17.51	−3.29	1998	.9073	14.05	−2.64
1977	.9996	−8.13	−.16	1999	.8939	20.83	1.09
1978	.9993	2.94	1.78	2000	.8945	−13.80	2.74
1979	.9978	15.88	2.46	2001	.8883	−18.39	−3.63
1980	.9977	22.03	.94	2002	.8822	−24.09	−2.23
1981	.9976	−18.26	−.04	2003	.8791	29.97	−.92
1982	.9977	13.64	4.23	2004	.8720	10.92	−.36
1983	.9969	11.96	−1.28	2005	.8686	4.72	.68

NOTE.—The table reports the fraction of total US mutual fund assets that are actively managed (S/W), the percentage return on that aggregate active portfolio, and the portfolio's market-adjusted return. The adjusted return is equal to the intercept plus the residual in a regression of the active portfolio's return on the market return. The value of S/W is for the end of the year during which the corresponding return occurs.

in D consist of $y_T = [r_{A,1} \cdots r_{A,T}]'$ and $z_T = \left[(S/W)_1 \cdots (S/W)_T\right]'$. In a regression of y_T on $-z_T$ and a constant, the intercept is a and the slope is b (see eq. [26]). Recall that investors' prior beliefs for a and b are given by the bivariate truncated normal distribution in equation (34), whose nontruncated moments are E_0 and V_0. Those moments are updated by using standard Bayesian results for the multiple regression model

$$V = \left[V_0^{-1} + \frac{1}{\sigma_x^2}(Z_T'Z_T)\right]^{-1}, \tag{36}$$

$$E = V\left(V_0^{-1}E_0 + \frac{1}{\sigma_x^2}Z_T'y_T\right), \tag{37}$$

where $Z_T = [\iota_T - z_T]$. The posterior distribution of a and b is bivariate truncated normal as in equation (34), except that E_0 and V_0 are replaced

by E and V from equations (36) and (37).[11] Having the updated moments E and V of the nontruncated bivariate normal distribution, we apply the relations in Muthen (1990) to obtain the updated moments of the truncated bivariate normal distribution, defined in equations (5) and (6).[12]

Panels C and D of figure 2 show the posterior distributions for a and b, respectively. Compared to the priors, the posteriors are shifted to the left, indicating a downward revision in beliefs about a and b. For example, for the middle prior (solid line), the posterior mean for a is 0.06, which is below the prior mean of 0.09, and the posterior mean for b is 0.07, below the prior mean of 0.1. Interestingly, while the posteriors are naturally tighter than the priors, they remain quite disperse. For example, for the middle prior, the values of zero and 0.15 are both well within the support of the posterior distributions of both a and b. Investors clearly remain substantially uncertain about a and b even after observing 44 years of data.

To understand why the posterior uncertainty about a and b is so large, it helps to examine figure 3. The figure plots the observed active fund returns $r_{A,t}$ against the observed active allocations $(S/W)_t$ for the full 44-year sample. The sample estimates of a and b can be obtained by fitting a line through the scatter plot in figure 3 and measuring the line's intercept ($= a$) and slope ($= -b$). It is immediately clear from figure 3 that these estimates are very imprecise because more than half the observations are bunched on top of each other at the right-hand-side edge of the plot; specifically, the first 27 observations in our 44-year sample have $(S/W)_t > 0.99$. Indeed, the confidence intervals for the ordinary least squares sample estimates of a and b are very wide (the standard errors of both estimates exceed 0.09, whereas the estimates themselves are both within 0.03 of zero). Since the sample does not contain much information about a and b, the posterior distributions in figure 2 are only modestly tighter than the priors.

The investor's beliefs about a and b, which are the key parameters in our model, translate into beliefs about α, a more familiar quantity. Recall that $\alpha = a - b(S/W)$ in the fully competitive setting. Figure 4 plots the prior and posterior distributions for the equilibrium α, which is α evaluated at the equilibrium value of S/W. The priors for α, shown in panel A, are computed from the priors of a and b and the same prior equilibrium values of $(S/W)_0$ as in figure 2: 0.6, 0.8, and 1.0. The poste-

[11] 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 one. As a result, the likelihood function is identical to what would arise if the observations of S/W were treated as nonstochastic.

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

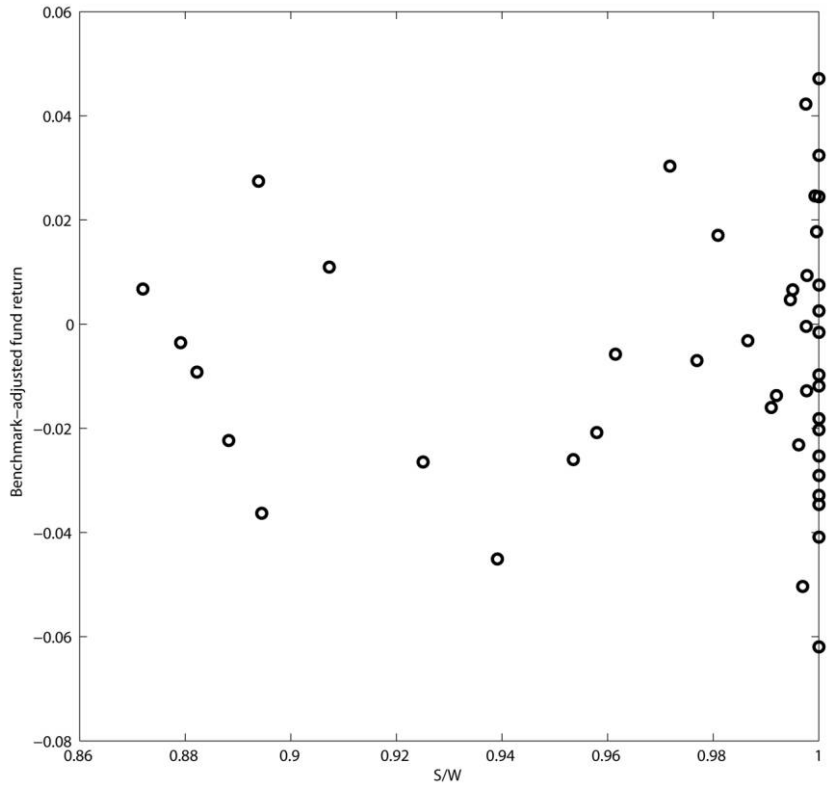

FIG. 3.—Active management's relative size (S/W) versus the benchmark-adjusted active return. The figure plots the annual observations of the fraction of total US mutual fund assets that are actively managed (horizontal axis) versus the market-adjusted percentage return on that aggregate active portfolio (vertical axis). The sample period is 1962–2005, and each S/W value is for the beginning of the year during which the corresponding return occurs.

riors for α, shown in panel *B*, are computed from the posteriors of a and b and the values of S/W that obtain in equilibrium on the basis of those posterior beliefs.

Figure 4 shows that all three priors are rather noninformative about α, in that most of the probability mass is on values between roughly −20 percent and 20 percent per year. The prior standard deviations range from 5.3 percent to 9.1 percent per year across the three priors. In contrast, the posteriors for α are much tighter: all three posterior standard deviations are just below 0.5 percent per year. The large difference between panels *A* and *B* indicates that our 44-year sample contains a lot of information about the equilibrium α. Another interesting comparison is that between the tight posteriors of α in figure 4 and the relatively disperse posteriors of a and b in figure 2. The large difference arises because a and b exhibit a very high posterior correlation (about 99 percent). In other words, after

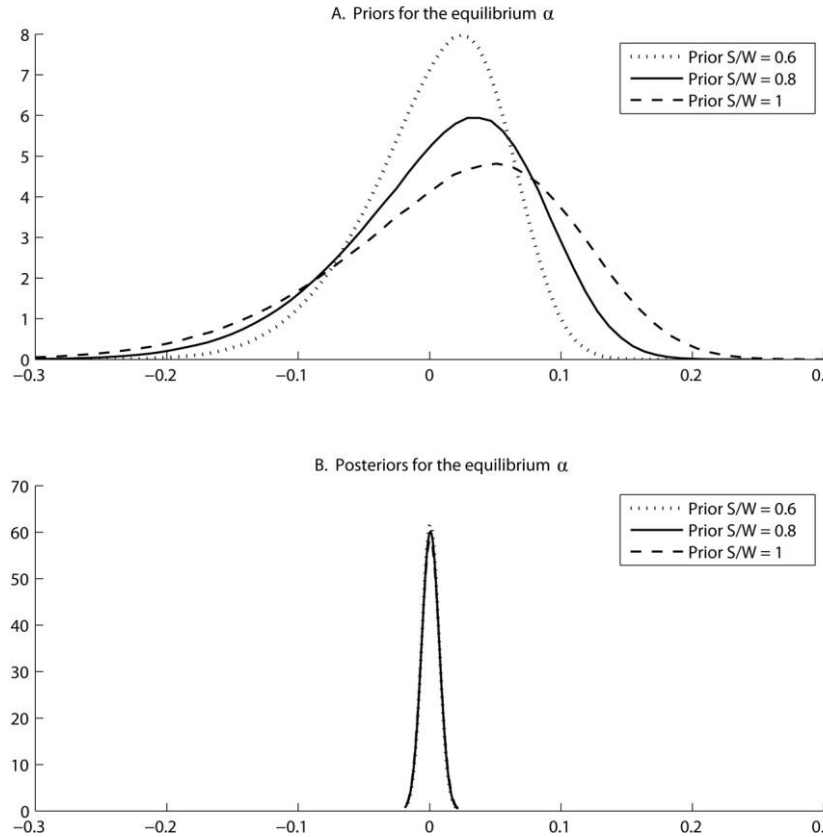

FIG. 4.—Prior and posterior distributions for alpha. The figure displays priors (panel *A*) and posteriors (panel *B*) of alpha in the fully competitive setting, where $\alpha_t = a - b(S/W)_t$, and $(S/W)_t$ is the aggregate allocation to active management. The priors and posteriors of α displayed here are implied by the priors and posteriors of a and b displayed in figure 2. The prior for b has a mean of 0.1; the three priors for a, based on that value, all have the property prob$(a < 0) = .01$ but differ with respect to the "prior S/W": the S/W under the fully competitive equilibrium based only on prior beliefs. The posteriors are obtained by updating the priors with the mutual fund data covering the 1962–2005 period.

observing the full sample, investors remain quite uncertain about the values of a and b, but they are quite certain that if a is high then b is high as well. They are also quite certain that the equilibrium α is close to zero. The posterior means of the equilibrium α are only about 7 basis points per year for all three priors, indicating that investors do not require much compensation for bearing the risk associated with active investing (see eq. [33]).

As a point of comparison, we also consider a dogmatic prior belief that returns to scale are constant ($b = 0$). Under that prior, only the beliefs about a are updated, following the standard result for updating the mean

of a normal distribution. Given the history of returns, $y_T = [r_{A,1} \ldots r_{A,T}]'$ with sample average $\bar{r}_{A,T}$, the posterior moments of a (and α) are given by

$$\tilde{\alpha} = \tilde{a} = \left(\frac{1}{V_0^{aa}} + \frac{T}{\sigma_x^2}\right)^{-1} \left(\frac{E_0^a}{V_0^{aa}} + \frac{T\bar{r}_{A,T}}{\sigma_x^2}\right), \tag{38}$$

$$\sigma_\alpha^2 = \sigma_a^2 = \left(\frac{1}{V_0^{aa}} + \frac{T}{\sigma_x^2}\right)^{-1}. \tag{39}$$

IV. Is the Industry's Size Puzzling Given Its Track Record?

In this section, we use our model to ask whether it is puzzling that the active management industry remains large, given its unflattering historical performance. Specifically, we explore properties of prior beliefs that would lead the rational investors in our model to choose a large allocation to active management, once they update their beliefs with the historical data. As discussed earlier, our model is not limited to mutual funds, but our empirical analysis treats that segment as representative of active management's relative size and track record, given the availability of mutual fund data. We thus take the most recent value of the S/W series described earlier, 0.87, as the empirical benchmark against which to compare equilibrium values of S/W implied by the model.

A. *Importance of Decreasing Returns to Scale*

Key to our analysis is the prior belief about b, the degree of decreasing returns to scale. The prior for b is fully determined by the prior mean b_0, as explained in Section III. Figure 5 plots the equilibrium allocation S/W for a wide range of values of b_0. For each value of b_0, the prior for the other unknown parameter, a, is specified such that prob$(a < 0) = .01$ and the equilibrium S/W based on just the prior beliefs about a and b is equal to 0.8 in the fully competitive setting $(M \to \infty, N \to \infty)$.[13] The values of S/W plotted in the figure are computed after updating beliefs about a and b with the histories of $r_{A,t}$ and $(S/W)_t$ used in the previous section.

Our main result appears as the solid line in figure 5, which plots the equilibrium S/W in the fully competitive setting. We see that the current size of the active management industry is consistent with beliefs that the industry faces decreasing returns to scale. For $b_0 \geq 0.1$ or so, the equilibrium S/W essentially matches the empirical benchmark value of 0.87. In-

[13] Our results are robust to alternative prior specifications, as discussed later in Sec. IV.D.

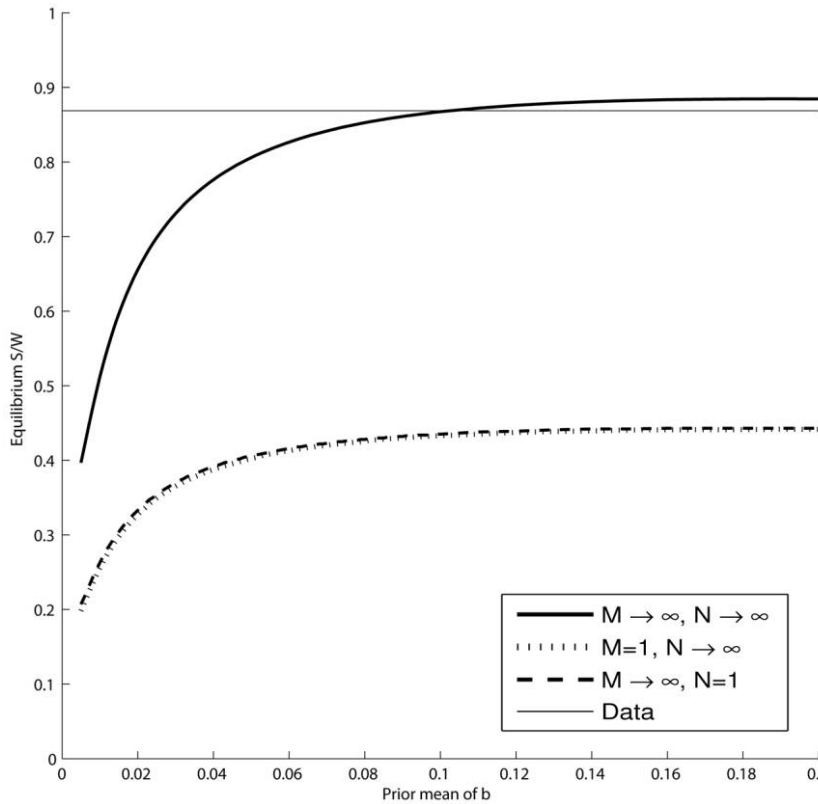

FIG. 5.—Equilibrium active allocation: effects of competition. For different numbers of funds (M) and investors (N), the figure plots the equilibrium aggregate allocation to active management (S/W) based on updated beliefs that incorporate mutual fund data for the 1962–2005 period. For each value of the prior mean of b, the prior for a is specified such that prob($a < 0$) = .01 and the equilibrium S/W equals 0.8 in the perfectly competitive case ($M \to \infty$, $N \to \infty$) when based only on the prior distribution for a and b. The "Data" line represents $S/W = 0.87$, the value at the end of 2005 for the mutual fund industry.

vestors are willing to invest that much 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. As discussed in the previous section, historical returns were earned at various levels of S/W all higher than 0.87, allowing investors to believe that performance going forward will be positive at $S/W = 0.87$.

With decreasing returns, investors' allocation to active management after observing poor performance can actually be higher than what they would allocate without seeing that performance. For $b_0 \geq 0.05$ or so, investors who allocate 80 percent to active management before seeing the data

allocate a higher fraction after seeing the data, even though the data include a negative return history. One reason is resolution of uncertainty about alpha. Recall that the posterior uncertainty about the equilibrium alpha is substantially lower than the prior uncertainty (fig. 4) and that this uncertainty (σ_α^2) appears in the denominator of S/W in (32). Another reason is that investors update their beliefs using not only the return history but also the history of $(S/W)_t$. A value of $b_0 = 0.1$ means that, before seeing the data, investors expect that reducing S/W by 0.1 would raise alpha by 1 percent. Recall from figure 2 that investors' posterior beliefs about b are shifted to the left relative to the prior. Although the posterior of a also shifts closer to zero, that shift is more than offset by the reductions in σ_α^2 and the mean of b, producing a higher ratio in (32).

The story is very different if investors believe that $b = 0$, that is, that returns to scale are constant. For example, if such investors have the same prior for a as in the case in which $b_0 = 0.1$, they invest nothing in active management after updating with the same historical data.[14] We also see that S/W in figure 5 takes substantially lower values as the prior mean of b moves closer to zero.

B. *Decreasing Returns versus Optimism*

The $b = 0$ setting also underscores the point that our story hinges on decreasing returns to scale ($b > 0$) as opposed to investor optimism about active management. Panel *A* of figure 6 displays percentiles of the prior for α in the $b = 0$ specification described above. This prior is the same as the distribution for a displayed as the solid line in panel *A* of figure 2. With $b = 0$, this is the prior for α at all levels of S/W. The corresponding prior for α in the $b > 0$ setting, with $b_0 = 0.1$, is displayed in panel *B* of figure 6. In the latter setting, the prior for α depends on S/W. For $S/W = 0$, the prior for α is the same as in the $b = 0$ case in panel *A*, but as S/W increases, all of the percentiles decline. In other words, at all positive levels of S/W, the prior for α is more optimistic when $b = 0$ than when $b > 0$. Despite the higher prior optimism about α, the equilibrium value of S/W equals zero after the $b = 0$ prior is updated with the data. In contrast, when the same data are used to update the less optimistic $b > 0$ prior, the resulting equilibrium S/W matches the empirical benchmark of 0.87. This striking difference arises because the $b = 0$ and $b > 0$ priors are updated very differently, as explained earlier in Section III.

[14] When $b = 0$, the cubic equation in (24) simplifies to a linear equation. In this case, σ_α^2 does not depend on S/W, so the equilibrium S/W is given directly by the first equality in (31). The active-management allocation problem is then essentially equivalent to the setting in Treynor and Black (1973) but with the addition of parameter uncertainty.

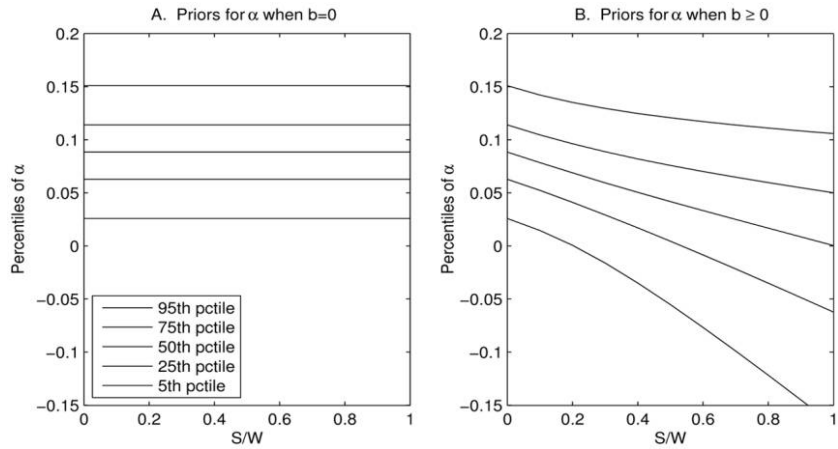

FIG. 6.—Priors for alpha: constant versus decreasing returns to scale. For each value of the aggregate allocation to active management (S/W), the figure displays percentiles of the prior distribution for alpha under constant returns to scale (panel *A*) and decreasing returns to scale (panel *B*), where $\alpha_t = a - b(S/W)_t$. The prior for a is displayed as the solid line in panel *A* of figure 2. Under constant returns to scale, the prior for b is dogmatic at $b = 0$. Under decreasing returns to scale, the prior for b has a mean of 0.1 and is displayed in panel *B* of figure 2. The ordering of the five lines in the plots is the same as in the legend box.

C. *Importance of Competition*

Competition plays an important role in determining the industry's size. Figure 5 also displays the equilibrium S/W when there is no competition among investors or no competition among managers. In analyzing these additional cases, we assume that investors have the same priors about a and b as in the fully competitive case, and they update with the same data. We see from the dashed line that solving the active-management allocation problem from the perspective of a single representative investor understates industry size by about half, even though managers compete with each other $(M \to \infty, N = 1)$. The dotted line shows that the industry's size is similarly understated by half if competitive investors allocate to an industry that acts as a monopolist $(M = 1, N \to \infty)$. Thus, we see that the effects of competition, obtained numerically in this mean-variance setting, follow closely the closed-form results obtained in the risk-neutral setting.

D. *Robustness*

Our results are quite robust to alternative specifications. Recall that for a given prior distribution of b, the prior for a is specified such that $\text{prob}(a < 0) = .01$ and S/W equals 0.8 on the basis of the prior for a and b. We now consider alternative values for these prior criteria. Specifically, we allow the prior value of S/W to be 1.0 or 0.6, and we allow $\text{prob}(a < 0)$ to be .1 or .001. Figure 7 displays the equilibrium S/W in the

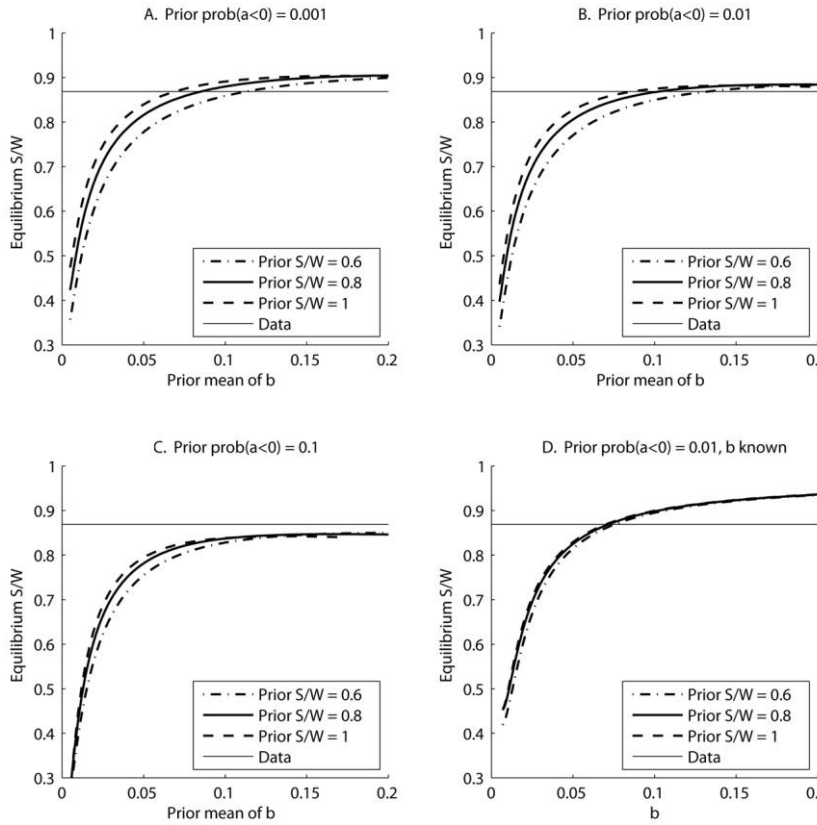

FIG. 7.—Equilibrium active allocation: robustness to priors. For alternative specifications of priors, the figure plots the fully competitive equilibrium aggregate allocation to active management (S/W), which appears as the solid curve in figure 5 under the original prior specification. Panel B maintains the original specification prob$(a < 0) = .01$ and considers three values—0.6, 0.8, and 1.0—for the "prior S/W," which is the equilibrium allocation based only on prior beliefs. The same three prior S/W values are then used with each of two alternative values of prob$(a < 0)$: panel A uses prob$(a < 0) = .001$, and panel C uses prob$(a < 0) = .1$. Panel D treats the case in which b is known with certainty to equal the value on the horizontal axis; this case specifies prob$(a < 0) = .01$ and considers the same three prior S/W values.

fully competitive setting after updating these alternative priors with the historical data. Each panel reports results under all three prior values of S/W. Panel B maintains the original specification of prob$(a < 0) = .01$; in panel A that probability is 10 times smaller, while in panel C it is 10 times larger.

While there are some differences across panels A–C of figure 7, all the results paint the same basic picture as the original result (solid line in fig. 5). That is, as b_0 increases, the equilibrium level of S/W rises sharply to a level approximately equal to the empirical benchmark. It does not rise quite

that high in panel *C*, which assigns a nontrivial 10 percent probability to even the first dollar of active management being unprofitable, but in general the results are robust to the various alternative specifications of the prior for a.

Panel *D* of figure 7 displays results in the hypothetical scenario in which the value of b is known a priori. The prior for a is specified using the same criteria as in the baseline case in panel *B*. Panel *D* shows that the empirical benchmark level of S/W is reached when b is known to be about 0.07, whereas reaching the same benchmark with unknown b requires a prior mean for b of about 0.1 (fig. 5). Removing uncertainty about b thus makes it easier to explain the empirical value of S/W with our decreasing-returns story, in that a known degree of decreasing returns can be weaker than what must be expected a priori when b is unknown.

Another result of knowing b is that the prior for a becomes less important. This result is demonstrated by the similarity of the three lines representing different prior S/W values in panel *D*. It also follows from the plots (not shown) based on the other two values of $\text{prob}(a < 0)$, which are virtually indistinguishable from panel *D*. The reason behind this result is that when b is known, the historical data become more informative about the single remaining unknown parameter, a.

V. Future Size of the Active Management Industry

While the previous section focuses on the present size of the active management industry, this section looks into the future. In Section V.A, we conduct simulations to calculate the expected future industry size after observing various potential track records. In Section V.B, we use the same simulations to investigate the speed of learning about returns to scale. To preview our results, we find that the industry is likely to remain large even if it continues to underperform. We also find that learning about a and b is slow, highlighting the rationale for treating these quantities as uncertain.

A. *Expected Future Industry Size*

In this subsection, we assess the expected future size of the active management industry conditional on a summary measure of the industry's future track record. We assume that investors enter the future with beliefs about a and b consistent with the industry's current size. We then compute the expected future values of the equilibrium S/W for different values of the t-statistic of the industry's future estimated alpha. We find that S/W is likely to remain large for a long period of time, even if the industry's future alpha turns out to be significantly negative as measured by the t-statistic.

We simulate 300,000 samples of future returns on active funds. To simulate a given sample, we first draw a and b randomly from their joint

posterior distribution at the end of our 44-year sample. We pick the baseline posterior distribution whose marginals are plotted by the solid lines in panels *C* and *D* of figure 2. Recall that this posterior is obtained from the prior for which $b_0 = 0.1$, the prior probability of $a < 0$ is 1 percent, and the prior equilibrium S/W is 0.8; the corresponding posterior equilibrium S/W of 0.867 approximately matches the observed value (see fig. 5). In the second step, we draw the random values of $x_t \sim N(0, \sigma_x^2)$ for $t = 1, \ldots, 20$ years. We construct the first future benchmark-adjusted active return as $r_{A,1} = a - b(0.867) + x_1$, following equation (26). On the basis of this return, we update the beliefs about a and b, following equations (36) and (37). We then solve for the new equilibrium allocation $(S/W)_2$ on the basis of those updated beliefs using proposition 2. Next, we construct $r_{A,2} = a - b(S/W)_2 + x_2$ and repeat the above procedure, building up the time series of $r_{A,t}$ and $(S/W)_t$ for $t = 1, \ldots, 20$. Note that $(S/W)_t$ affects $r_{A,t}$, which in turn affects $(S/W)_{t+1}$, and so on. For every t, we compute an estimate of the industry's alpha, or $\hat{\alpha}$, as the sample average of $\{r_{A,1}, \ldots, r_{A,t}\}$, and we calculate the t-statistic $\hat{\alpha}\sqrt{t}/\sigma_x$. Finally, we compute the expected $(S/W)_t$ conditional on a given value t_0 of the t-statistic as the average value of $(S/W)_t$ across all simulated samples producing t-statistics within a small neighborhood of t_0. The results are plotted in panel *A* of figure 8.

Panel *A* of figure 8 plots the expected future values of $(S/W)_t$, or $\mathrm{E}(S/W)$, conditional on the future t-statistics of -2, 0, and 2. Conditional on the t-statistic of 0, $\mathrm{E}(S/W)$ is roughly constant, declining from the current value of 86.7 percent to 86.2 percent after 20 years. The slight decline reflects the mild disappointment of investors who expect to earn a slightly positive $\hat{\alpha}$ (of about 7 basis points per year, as noted in the description of fig. 4) but end up earning $\hat{\alpha} = 0$. Conditional on the t-statistic of -2, $\mathrm{E}(S/W)$ declines over time. Not surprisingly, if the industry's performance turns out to be worse than expected, the industry is expected to shrink to the benefit of passive investments. More interesting, $\mathrm{E}(S/W)$ remains substantial for long periods of time: it declines from 86.7 percent to 72.5 percent after 10 years and to 63.4 percent after 20 years. That is, the industry is expected to remain large even if it continues to significantly underperform its benchmark. This striking result is due to decreasing returns to scale. Investors observing underperformance reduce their active allocation, but not as much as they would under constant returns to scale because they understand that when they allocate less to active management, their future active returns will be higher. (Specifically, investors reduce S/W until α reaches its positive equilibrium level in eq. [33].)

Panel *A* also shows that conditional on the t-statistic of 2, $\mathrm{E}(S/W)$ rises to 92.2 percent after 20 years. If the industry performs better than expected, it grows at the expense of passive investments, but its growth is restrained by decreasing returns to scale: investors know that when they al-

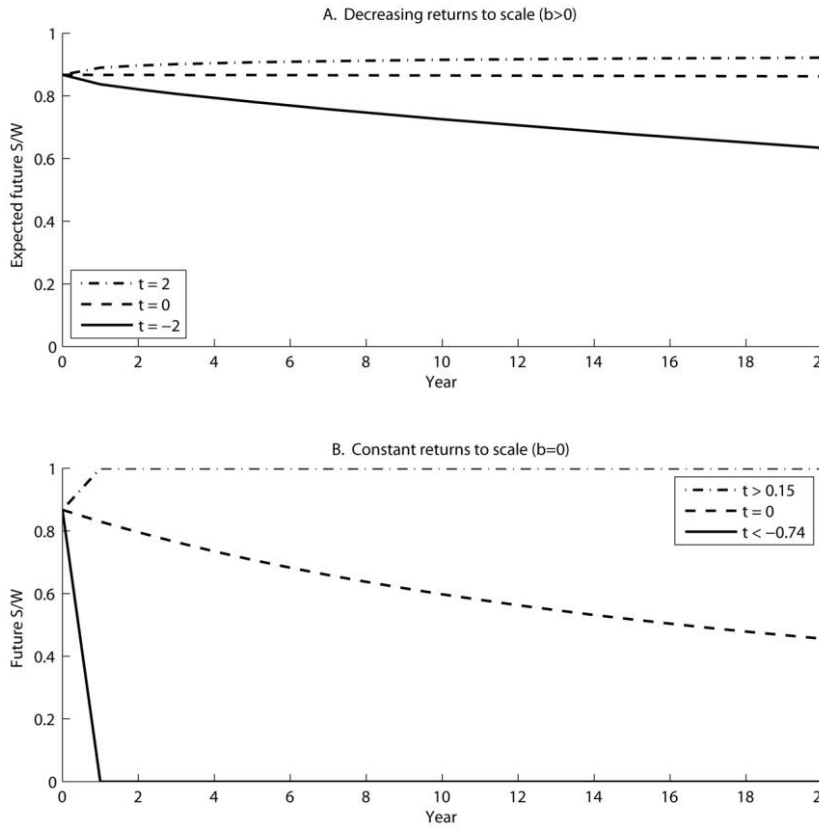

FIG. 8.—Expected future industry size conditional on given future performance. This figure plots the expected values of S/W, the fully competitive equilibrium aggregate allocation to active management, over 20 future years for different t-statistics of the future sample estimate of the active industry's alpha. Panel A plots the expected values of S/W conditional on future t-statistics of -2, 0, and 2 under decreasing returns to scale. In this case, future active returns are simulated on the basis of the posteriors of a and b represented by the solid lines in figure 2. Panel B plots the future values of S/W implied by three sets of future t-statistics, $t > 0.15$, $t = 0$, and $t < -0.74$, under constant returns to scale. In this case, $b = 0$ and the distribution of a is normal with the same mean and variance as the posterior distribution of α in figure 4.

locate more to active management, their future returns will be lower. Because of this key mechanism, S/W is expected to vary slowly over time regardless of performance. Overall, panel A of figure 8 shows that the active management industry is likely to remain large for many years.

A very different picture emerges if investors believe a priori that returns to scale are constant ($b = 0$). In that case, there is no need to simulate because the t-statistic of $\hat{\alpha}$ is a sufficient statistic for S/W: a given future t-statistic implies a unique future value of S/W. To produce a fair compar-

ison with the $b > 0$ case discussed above, we choose the distribution of a such that investors with the $b = 0$ prior perceive the same mean and variance of α as in the $b > 0$ case. Specifically, since $b = 0$ implies $a = \alpha$, we assume that the distribution of a is normal with the same mean and variance as the posterior distribution of α obtained under the $b > 0$ prior (see panel *B* of fig. 4). As a result, investors initially choose the same S/W (of 86.7 percent) in both cases (see eq. [31]). The resulting future values of S/W are plotted in panel *B* of figure 8.

Panel *B* of figure 8 shows that under constant returns to scale, the industry's size is much more sensitive to performance. Conditional on the *t*-statistic of 0, S/W drops from 86.7 percent to 45.6 percent after 20 years. The reason behind the drop is the same as in the $b > 0$ case, but the magnitude is much larger. When $b = 0$, the response of S/W to performance is no longer cushioned by decreasing returns to scale in that a reduction in S/W no longer implies a higher expected future return. The results are even more dramatic when we condition on nonzero *t*-statistics. For any *t*-statistic greater than 0.15, S/W jumps to one after the very first year. For any *t*-statistic below -0.74, we obtain the other corner solution, $S/W = 0$, after just 1 year of underperformance. These implications appear less plausible than those obtained under decreasing returns to scale.

In simulating the future, we apply our model's equilibrium in multiple successive years. As explained earlier, the active return r_A in each year depends on the equilibrium active allocation S/W chosen in the previous year. In that sense, our model delivers a year-by-year dependence between r_A and S/W, generally implying that an unexpectedly high r_A in a given year causes a higher S/W going into the next year. In principle, one could also look for this dependence in the year-by-year historical data in table 1, but we do not believe that such an exercise would be very informative. Indexing was novel when it emerged on the investment landscape during the 1970s. Understanding subsequent year-by-year fluctuations in its share relative to active management must surely have much to do with the dissemination and adoption of financial innovation, which we cannot hope to capture in our simple model. The strength and duration of the innovation-related effects in the historical year-by-year variation of S/W are difficult to assess, and we think that a conservative approach here is best. That is, we simply assume that indexing, by now, has evolved to a fairly mature and familiar alternative, enough so that it is reasonable to entertain current and future active versus passive decisions as being dependent on track records.

B. Learning about Returns to Scale

The values of a and b are unknown to investors. How fast can investors learn about a and b by observing realized returns and active allocations?

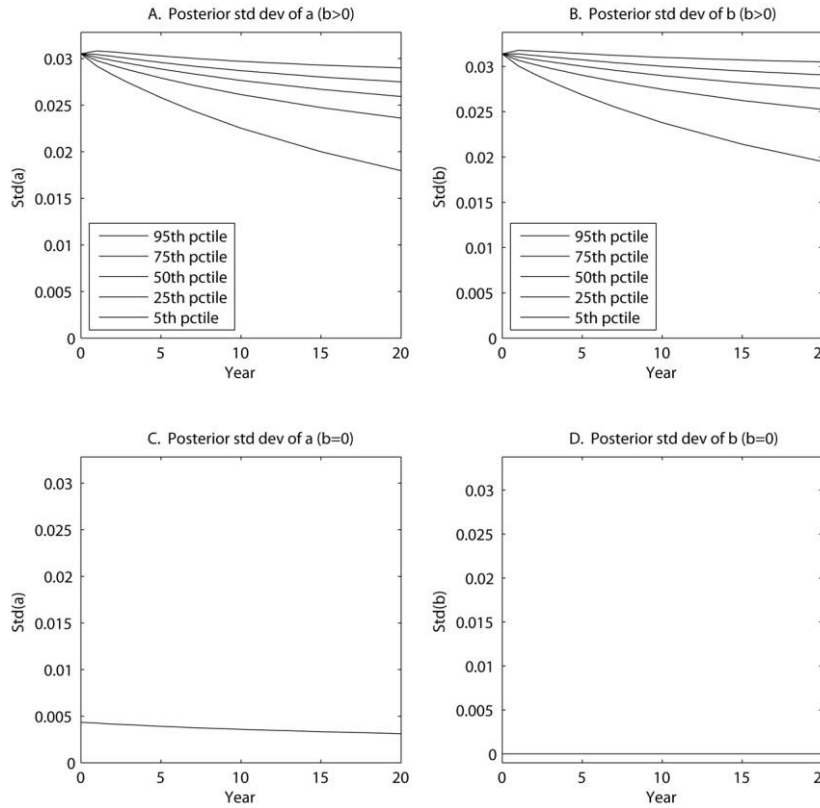

FIG. 9.—Speed of learning about a and b. Panels A and B plot the evolution of selected percentiles of the distributions of the posterior standard deviations of a and b, respectively, over 20 future years under decreasing returns to scale. In this case, future active returns are simulated on the basis of the posteriors of a and b represented by the solid lines in figure 2. Panel C plots the evolution of the posterior standard deviation of a over 20 future years under constant returns to scale. In this case, $b = 0$ and the distribution of a is normal with the same mean and variance as the posterior distribution of α in figure 4. Panel D simply indicates that the posterior standard deviation of b under constant returns to scale is zero. The ordering of the five lines in panels A and B is the same as in the legend box.

To answer this question, we rely on the 300,000 simulated samples described in the previous subsection.

As investors learn, their posterior standard deviations of a and b decline over time. The rate of decline in these standard deviations depends on the true values of a and b. We draw these values from their joint posterior distribution, as described earlier. The probability distribution of a and b thus gives rise to distributions of the posterior standard deviations of a and b. Panels A and B of figure 9 show the evolution of these distributions over time. Both panels plot selected percentiles of the distribu-

tion of the respective standard deviation across the 300,000 simulated samples.

Panels *A* and *B* of figure 9 show that learning about a and b is typically slow. For the median sample, the posterior standard deviations decline only modestly over the 20-year period: from 0.030 to 0.026 for a and from 0.031 to 0.028 for b. Investors thus typically remain highly uncertain about a and b even after 20 additional years of learning (on top of the 44 years in our sample). This result underscores the importance of incorporating uncertainty about a and b in assessing the size of the active management industry.

Why is learning about a and b slow? 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. As explained in Section V.A, fluctuations in S/W are muted by decreasing returns to scale. The resulting stability of S/W hampers learning about a and b. To see why, recall that a and b represent the intercept and slope from the regression of $r_{A,t}$ on $-(S/W)_t$. If the right-hand-side variable in the regression does not fluctuate much, learning about the intercept and slope is slow. At the extreme, if S/W stops fluctuating, learning about a and b stops as well. In that case, investors would eventually learn the true value of α at the prevailing level of S/W, but they would never learn a and b, so they would forever remain uncertain about α at any other level of S/W.

The extreme case in which S/W stops fluctuating also helps illustrate the link between slow learning and competition among investors. The aggregate active allocation S/W is determined in equilibrium by competing investors who cannot coordinate their investment decisions. If investors could instead coordinate, they might well find it useful to continue varying S/W 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 that would make 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.

The speed of learning about a and b can be faster or slower than in the median case discussed above, depending on the path of S/W. For example, the 95th percentile of the posterior standard deviation of a after 20 years is 0.029, which is only slightly smaller than the initial value of 0.030. For these simulated samples, in which S/W fluctuates the least, hardly any learning takes place. In contrast, the 5th percentile of the same distribution is only 0.018, indicating much faster learning for samples in

which S/W fluctuates more. This interesting path dependence of the speed of learning is a direct consequence of decreasing returns to scale.[15]

In contrast to the fascinating learning process under decreasing returns to scale, learning under constant returns to scale is straightforward (see eqq. [38] and [39]). With $b = 0$, the value of a ($= \alpha$) is simply the unconditional mean return. The posterior standard deviation of a, plotted in panel C of figure 9, declines at the usual $\sqrt{t}$ rate, regardless of the particular sample realization. Investors learn differently under decreasing returns to scale because the level and variation in $(S/W)_t$ affect learning when $b > 0$ but not when $b = 0$. To summarize, learning about decreasing returns to scale is path dependent and generally slow, leaving investors highly uncertain about a and b even after observing long histories of returns and active allocations.

VI. 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 the seminal work of Berk and Green (2004), but there are important differences. First, Berk and Green assume that decreasing returns apply at the level of individual funds, whereas we assume that they apply to the active management industry as a whole. That is, we assume that an individual fund's alpha is decreasing in the total amount invested by all active funds.[16] 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.[17] 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

[15] As an aside, panels A and B show that for a small subset of simulated samples, the posterior standard deviations exceed the prior ones in the first few years. This result is due to truncation in the prior for b ($b \geq 0$). The standard deviation of any left-truncated normal distribution is increasing in the mean of the same distribution. Therefore, sample evidence that raises the posterior mean also pushes up the posterior standard deviation. This force may be stronger or weaker than the offsetting effect of learning, which always pulls the posterior standard deviation down.

[16] 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 on returns to scale at the fund level for mutual funds is provided by Chen et al. (2004), Pollet and Wilson (2008), and Reuter and Zitzewitz (2011). Related evidence for hedge funds, at the fund level as well as aggregate level, is provided by Fung et al. (2008).

[17] A similar perspective is adopted by Glode and Green (2011), who argue that fund returns can be decreasing in the size of a sector or trading strategy, as well as in the size of the fund itself. Glode and Green develop a model of information spillovers that can rationalize performance persistence in hedge funds.

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 Berk and Green (2004) corresponds to a setting in which a is unknown but b is known.[18] As discussed earlier, when both a and b in (4) are unknown, investors face an interesting learning problem in which learning about those parameters is generally slow.

Another difference from Berk and Green (2004) is that their investors face $\tilde{\alpha} = 0$, whereas ours perceive $\tilde{\alpha} > 0$. Our investors maximize (9). Berk and Green do not solve the investors' optimization problem explicitly; instead, they fix $\tilde{\alpha} = 0$ by invoking the assumption that nonbenchmark risk can be completely diversified away across many funds. Berk and Green 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 nonbenchmark risk; $\tilde{\alpha} = 0$ is thus a necessary condition for equilibrium. Recall from proposition 1 that our model also implies $\tilde{\alpha} = 0$ in the special case of perfect competition with risk-neutral investors. If investors are risk averse, then $\tilde{\alpha} > 0$ because investors require compensation for both nondiversifiable risk (σ_x) and uncertainty about α (σ_α), as shown in equation (33). However, $\tilde{\alpha}$ in the competitive setting is not necessarily large, especially if learning proceeds to the point at which σ_α is small. For example, the posterior mean of α in our figure 4 is only 7 basis points per year, as noted earlier. Thus, even though our modeling of the determinants of equilibrium alpha is rather different from that of Berk and Green, their zero-alpha condition is not at sharp odds with our model in practical terms.[19]

VII. Conclusion

It seems puzzling that active management remains popular despite its poor track record. We propose a potential resolution to this puzzle. Using a model with competing investors and fund managers, we find that the large observed size of the active management industry can be rationalized if investors believe that active managers face decreasing returns to scale. If investors instead believed that returns to scale were constant, they would

[18] Berk and Green 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 eq. [2]).

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

allocate nothing to active management today, even if they were initially more optimistic about active managers' abilities.

Under decreasing returns to scale, investors adjust their allocation in response to performance to achieve the desired expected return going forward. After a period of underperformance, the proportional allocation to active management should be smaller than it was at the beginning of the period, but it should also remain substantial. Both predictions are consistent with the empirical evidence for active mutual funds, which have underperformed passive benchmarks over the past four decades: passive investing has grown dramatically since its humble beginnings in the 1970s, but active investing remains more popular to this day. We also show that the active management industry is likely to remain large for many more years, even if it continues to perform poorly.

Investors in our model face endogeneity that limits their learning: what they learn affects how much they allocate to active management, and what they allocate affects how much they learn. Owing to this endogeneity, the equilibrium allocation tends to vary little over time, resulting in slow learning about the degree of returns to scale in active management. Initial beliefs about returns to scale thus affect the investors' active allocations for a long time.

Given the inherent difficulty in estimating returns to scale, further empirical work seems warranted. Besides estimating returns to scale at the aggregate level, one could also try to measure them for various segments of the active management industry. Future research can also explore additional aspects of learning about the parameters governing returns to scale. Those parameters are held constant in our model for simplicity, but they could plausibly vary because of exogenous shocks, such as shocks to market liquidity. In such a setting, parameter uncertainty would get refreshed periodically, further slowing the learning process. Continuing research into decreasing returns to scale in active management is likely to yield nondecreasing returns.

Appendix

This appendix provides details of the equilibria in the various settings considered. After brief preliminaries, we first analyze the risk-neutral setting and then turn to the mean-variance setting. In both settings we consider three cases: perfect competition among managers and investors ($M \to \infty$, $N \to \infty$), perfectly competitive managers facing a single investor ($M \to \infty$, $N = 1$), and perfectly competitive investors facing a single manager ($M = 1$, $N \to \infty$).

Combining equations (8) and (11) gives investor j's excess portfolio return as

$$r_j = r_P + \delta_j' \left(a\iota_M - b \frac{S}{W} \iota_M - \underline{f} + u \right), \tag{A1}$$

where $\underline{f}$ is the $M \times 1$ vector of fund fees. Denote $\omega_j = \delta_j' \iota_M$. We then obtain

$$\mathrm{E}(r_j|D) = \mu_P + \left(\tilde{a} - \tilde{b}\frac{S}{W}\right)\omega_j - \delta_j'\underline{f}, \tag{A2}$$

$$\mathrm{Var}(r_j|D) = \sigma_P^2 + \left[\sigma_a^2 + \sigma_x^2 + \sigma_b^2\left(\frac{S}{W}\right)^2 - 2\sigma_{ab}\frac{S}{W}\right]\omega_j^2 + \sigma_\epsilon^2\left(\delta_j'\delta_j\right). \tag{A3}$$

When $\tilde{a} \leq 0$, investors invest nothing in active management (recall that the elements of δ_j must be nonnegative). In that case, a positive investment in active management would produce a lower expected return and higher variance than an all-benchmark investment; therefore, $S/W = 0$ in equilibrium. Since active management does exist, we assume hereafter that $\tilde{a} > 0$.

When managers are perfectly competitive ($M \to \infty$), it is clear that $\underline{f}$ must be the zero vector in equilibrium. Any manager charging a positive fee would be offering investors a lower expected return than any zero-fee competitors. In a risk-neutral setting, it follows immediately that such a manager would receive no investment. In a mean-variance setting, the presence of many competing managers allows investors to hold well-diversified portfolios, with the property that $\delta_j'\delta_j \to 0$, so the positive-fee manager offers no reduction in overall variance. Thus, compared to his many zero-fee competitors, a positive-fee manager would simply be offering a lower expected return with no reduction in variance, and he would again receive no investment.

A. *Risk-Neutral Setting*

$M \to \infty$ and $N \to \infty$: When investors are perfectly competitive ($N \to \infty$), investor j views the choice of ω_j as having no effect on S/W. Since $\underline{f}$ is zero with perfectly competitive managers, it follows from (A2) that each risk-neutral investor chooses ω_j to maximize the expected return

$$\mathrm{E}(r_j|D) = \mu_P + \left(\tilde{a} - \tilde{b}\frac{S}{W}\right)\omega_j. \tag{A4}$$

If (23) does not bind, investor j's first-order condition in maximizing (A4) is (15), which then delivers (14), using (8) and $f_i = 0$. If (23) binds, then every investor desires $\omega_j > 1$ and $S/W = 1$. In that case, $\tilde{\alpha} = \tilde{a} - \tilde{b}$ is positive.

$M \to \infty$ and $N = 1$: When $N = 1$, the single investor realizes that $\omega_j = S/W$ and replaces (A4) with

$$\mathrm{E}(r_j|D) = \mu_P + \left(\tilde{a} - \tilde{b}\frac{S}{W}\right)\frac{S}{W}. \tag{A5}$$

If (23) does not bind, the equilibrium value of S/W that maximizes expected return is given by (18). If (23) binds, then, as in the previous case, $S/W = 1$ and $\tilde{\alpha} = \tilde{a} - \tilde{b}$.

$M = 1$ and $N \to \infty$: Here the monopolistic manager sets the rate f to maximize fee revenue fS. For a given f, we see from (A2) that each investor j chooses ω_j to maximize

$$\mathrm{E}(r_j|D) = \mu_P + \left(\tilde{a} - \tilde{b}\frac{S}{W} - f\right)\omega_j, \tag{A6}$$

giving the first-order condition

$$S = W\frac{\tilde{a} - f}{\tilde{b}}, \tag{A7}$$

and thus

$$fS = W\frac{f(\tilde{a} - f)}{\tilde{b}}. \tag{A8}$$

Knowing (A8), the manager sets the maximizing value $f = \tilde{a}/2$, as given in (19). Substituting that value into (A7) implies $S/W = \tilde{a}/(2\tilde{b})$, as given in (21). Substituting those values for f and S/W into (8) gives (20). If $\tilde{a}/(2\tilde{b}) > 1$, then satisfying (23) requires $S/W = 1$ and, therefore, $f = \tilde{a} - \tilde{b}$.

B. *Mean-Variance Setting*

$M \to \infty$ and $N \to \infty$: Since in this setting we can set $\underline{f}$ and $\delta_j'\delta_j$ to zero, as discussed earlier, each investor solves

$$\max_{\omega_j}\left\{\frac{\mathrm{E}(r_j|D)}{\sqrt{\mathrm{Var}(r_j|D)}}\right\}, \tag{A9}$$

where $\mathrm{E}(r_j|D)$ is given by (A4) and $\mathrm{Var}(r_j|D)$ is given by

$$\mathrm{Var}(r_j|D) = \sigma_P^2 + \left[\sigma_a^2 + \sigma_x^2 + \sigma_b^2\left(\frac{S}{W}\right)^2 - 2\sigma_{ab}\frac{S}{W}\right]\omega_j^2. \tag{A10}$$

When the constraint in (23) does not bind, the first-order condition for the maximization in (A9) is

$$0 = \left(\tilde{a} - \tilde{b}\frac{S}{W}\right)\sigma_P^2 - \omega_j\left[\sigma_a^2 + \sigma_x^2 + \sigma_b^2\left(\frac{S}{W}\right)^2 - 2\sigma_{ab}\frac{S}{W}\right]\mu_P. \tag{A11}$$

Dividing through by σ_P^2, recalling $\gamma = \mu_P/\sigma_P^2$, and recognizing that $\omega_j = S/W$ in equilibrium gives the cubic equation in (24). It can be verified that this equation has one positive real solution for S/W. If that solution exceeds one, the constraint in (23) binds, and then $S/W = 1$.

$M \to \infty$ and $N = 1$: As in the above case, we can set $\underline{f}$ and $\delta_j'\delta_j$ to zero. As in the earlier risk-neutral setting, a single investor solving (A9) realizes that $\omega_j = S/W$. The expected return is then given by (A5), and the variance is given by

$$\mathrm{Var}(r_j|D) = \sigma_P^2 + \left(\frac{S}{W}\right)^2(\sigma_a^2 + \sigma_x^2) + \left(\frac{S}{W}\right)^4\sigma_b^2 - 2\left(\frac{S}{W}\right)^3\sigma_{ab}. \tag{A12}$$

Equilibrium is computed by using (A5) and (A12) to solve (A9) numerically, subject to the constraint $S/W \leq 1$.

$M = 1$ and $N \to \infty$: As in the risk-neutral setting, the expected return is given by (A6). In this single-manager case, we consider the aggregate portfolio of active funds as if it were managed by a monopolist, so $\sigma_\epsilon = 0$ for this diversified portfolio. The investor's return variance is then given by (A3). Substituting the equilibrium condition $\omega_j = S/W$ into investor j's first-order condition for the maximization in (A9) leads to the cubic equation

$$0 = \tilde{a} - f - \frac{S}{W}\left[\tilde{b} + \gamma\left(\sigma_a^2 + \sigma_x^2\right)\right] + \left(\frac{S}{W}\right)^2 2\gamma\sigma_{ab} - \left(\frac{S}{W}\right)^3 \gamma\sigma_b^2, \tag{A13}$$

which is the same as (24) but with $\tilde{a}$ replaced by $\tilde{a} - f$. Equilibrium is computed numerically by finding the value of f that maximizes f times the solution to (A13) given f.

References

Avramov, Doron, and Russ Wermers. 2006. "Investing in Mutual Funds When Returns Are Predictable." *J. Financial Econ.* 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." *J. Finance* 56:45–85.

Berk, Jonathan B., and Richard C. Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *J.P.E.* 112:1269–95.

Bolton, Patrick, Tano Santos, and José A. Scheinkman. 2011. "Cream Skimming in Financial Markets." Manuscript, Columbia Univ.

Chen, Joseph, Harrison Hong, Ming Huang, and Jeffrey Kubik. 2004. "Does Fund Size Erode Mutual Fund Performance?" *A.E.R.* 94:1276–1302.

Chordia, Tarun. 1996. "The Structure of Mutual Fund Charges." *J. Financial Econ.* 41:3–39.

Cuoco, Domenico, and Ron Kaniel. 2011. "Equilibrium Prices in the Presence of Delegated Portfolio Management." *J. Financial Econ.* 101:264–69.

Dangl, Thomas, Youchang Wu, and Josef Zechner. 2008. "Market Discipline and Internal Governance in the Mutual Fund Industry." *Rev. Financial Studies* 21:2307–43.

Das, Sanjiv R., and Rangarajan K. Sundaram. 2002. "Fee Speech: Signaling, Risk-Sharing, and the Impact of Fee Structures on Investor Welfare." *Rev. Financial Studies* 15:1465–97.

Dasgupta, Amil, Andrea Prat, and Michela Verardo. 2011. "The Price Impact of Institutional Herding." *Rev. Financial Studies* 24:892–925.

Del Guercio, Diane, and Jonathan Reuter. 2011. "Mutual Fund Performance and the Incentive to Invest in Active Management." Manuscript, Univ. Oregon.

Fama, Eugene F., and Kenneth R. French. 2007. "Disagreement, Tastes, and Asset Prices." *J. Financial Econ.* 83:667–89.

———. 2010. "Luck versus Skill in the Cross Section of Mutual Fund Returns." *J. Finance* 65:1915–47.

French, Kenneth R. 2008. "Presidential Address: The Cost of Active Investing." *J. Finance* 63:1537–73.

Fung, William, David A. Hsieh, Narayan Y. Naik, and Tarun Ramadorai. 2008. "Hedge Funds: Performance, Risk, and Capital Formation." *J. Finance* 63:1777–1803.
Garcia, Diego, and Joel M. Vanden. 2009. "Information Acquisition and Mutual Funds." *J. Econ. Theory* 144:1965–95.
Glode, Vincent. 2011. "Why Mutual Funds 'Underperform.'" *J. Financial Econ.* 99:546–59.
Glode, Vincent, and Richard C. Green. 2011. "Information Spillovers and Performance Persistence for Hedge Funds." *J. Financial Econ.* 101:1–17.
Grossman, Sanford G., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *A.E.R.* 70:393–408.
Gruber, Martin J. 1996. "Another Puzzle: The Growth in Actively Managed Mutual Funds." *J. Finance* 51:783–810.
Guerrieri, Veronica, and Peter Kondor. 2012. "Fund Managers, Career Concerns, and Asset Price Volatility." *A.E.R.* 102 (5): 1986–2017.
He, Zhiguo, and Arvind Krishnamurthy. Forthcoming. "Intermediary Asset Pricing." *A.E.R.*
Huang, Jennifer, Kelsey D. Wei, and Hong Yan. 2007. "Participation Costs and the Sensitivity of Fund Flows to Past Performance." *J. Finance* 62:1273–1311.
Investment Company Institute. 2009. *Investment Company Fact Book.* Washington, DC: Investment Co. Inst.
Jensen, Michael C. 1968. "The Performance of Mutual Funds in the Period 1945–1964." *J. Finance* 23:389–416.
Khorana, Ajay, Henri Servaes, and Peter Tufano. 2005. "Explaining the Size of the Mutual Fund Industry around the World." *J. Financial Econ.* 78:145–85.
Kogan, Leonid, Stephen A. Ross, Jiang Wang, and Mark M. Westerfield. 2006. "The Price Impact and Survival of Irrational Traders." *J. Finance* 61:195–229.
Lynch, Anthony W., and David K. Musto. 2003. "How Investors Interpret Past Returns." *J. Finance* 58:2033–58.
Malkiel, Burton G. 1995. "Returns from Investing in Equity Mutual Funds 1971 to 1991." *J. Finance* 50:549–72.
Mamaysky, Harry, and Matthew Spiegel. 2002. "A Theory of Mutual Funds: Optimal Fund Objectives and Industry Organization." Manuscript, Yale Univ.
Muthen, Bengt. 1990. "Moments of the Censored and Truncated Bivariate Normal Distribution." *British J. Math. and Statis. Psychology* 43:131–43.
Nanda, Vikram, M. P. Narayanan, and Vincent A. Warther. 2000. "Liquidity, Investment Ability, and Mutual Fund Structure." *J. Financial Econ.* 57:417–43.
Pástor, Ľuboš, and Robert F. Stambaugh. 2002*a*. "Investing in Equity Mutual Funds." *J. Financial Econ.* 63:351–80.
———. 2002*b*. "Mutual Fund Performance and Seemingly Unrelated Assets." *J. Financial Econ.* 63:315–49.
Petajisto, Antti. 2009. "Why Do Demand Curves for Stocks Slope Down?" *J. Financial and Quantitative Analysis* 44:1013–44.
Philippon, Thomas. 2008. "The Evolution of the U.S. Financial Industry from 1860 to 2007: Theory and Evidence." Manuscript, New York Univ.
Pollet, Joshua, and Mungo Wilson. 2008. "How Does Size Affect Mutual Fund Behavior?" *J. Finance* 63:2941–69.
Reuter, Jonathan, and Eric Zitzewitz. 2011. "How Much Does Size Erode Mutual Fund Performance? A Regression Discontinuity Approach." Manuscript, Boston Coll.
Rosenbaum, S. 1961. "Moments of a Truncated Bivariate Normal Distribution." *J. Royal Statis. Soc., Ser. B (Methodological)*, 21:405–8.

Savov, Alexi. 2009. "Free for a Fee: The Hidden Cost of Index Fund Investing." Manuscript, New York Univ.
Sharpe, William F. 1991. "The Arithmetic of Active Management." *Financial Analysts J.* 47 (January/February): 7–9.
Stein, Jeremy C. 2005. "Why Are Most Funds Open-End? Competition and the Limits of Arbitrage." *Q.J.E.* 120:247–72.
Treynor, Jack L., and Fischer Black. 1973. "How to Use Security Analysis to Improve Portfolio Selection." *J. Bus.* 46:66–86.
Vayanos, Dimitri, and Paul Woolley. 2008. "An Institutional Theory of Momentum and Reversal." Manuscript, London School Econ.
Wermers, Russ. 2000. "Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses." *J. Finance* 55:1655–95.