Active and Passive Investing

Nicolae Gârleanu and Lasse Heje Pedersen*

October 1, 2019

Abstract

We model how investors allocate between asset managers, managers choose their portfolios of multiple securities, fees are set, and security prices are determined. The optimal passive portfolio is linked to the "expected market portfolio," while the optimal active portfolio has elements of value and quality investing. We make precise Samuelson's Dictum by showing that macro inefficiency is greater than micro inefficiency under realistic conditions — in fact, all inefficiency arises from systematic factors when the number of assets is large. Further, we show how the costs of active and passive investing affect macro and micro efficiency, fees, and assets managed by active and passive managers. Our findings help explain empirical facts about the rise of delegated asset management, the composition of passive indices, and the resulting changes in financial markets.

Keywords: asset pricing, market efficiency, asset management, search, information
JEL Codes: D4, D53, D8, G02, G12, G14, G23, L1

*Gârleanu is at the Haas School of Business, University of California, Berkeley, and NBER; e-mail: garleanu@berkeley.edu. Pedersen is at AQR Capital Management, Copenhagen Business School, New York University, and CEPR; www.lhpedersen.com. We are grateful for helpful comments from Antti Ilmanen, Kelvin Lee, and Peter Norman Sørensen as well as from seminar participants at the Berkeley-Columbia Meeting in Engineering and Statistics, Federal Reserve Bank of New York, and Copenhagen Business School. Pedersen gratefully acknowledges support from the FRIC Center for Financial Frictions (grant no. DNRF102). AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR.

Over the past half century, financial markets have witnessed a continual rise of delegated asset management and, especially over the past decade, a marked rise of passive management, as seen in Figure 1. This delegation has potentially profound implications for market efficiency (see, e.g., the presidential addresses to the American Finance Association of Grossman (1995), Stein (2009), and Stambaugh (2014)), investor behavior (presidential address of Gruber (1996)), and asset management fees (e.g., the presidential address of French (2008)).

The rise of delegated management raises several questions: What is the optimal portfolio of active (i.e., informed) and passive (i.e., uninformed) managers, respectively? What determines the number of investors choosing active management, passive management, or direct holdings? What are the implications of delegated management on market efficiency at the micro and macro levels? How do macro and micro efficiencies depend on the costs of active and passive management?

We address these questions in an asymmetric-information equilibrium model where security prices, asset management fees, portfolio decisions, and investor behavior are jointly determined. Our main findings are: (1) the optimal passive portfolio is the "expected market portfolio," tilted away from assets with the most supply uncertainty; (2) the optimal active portfolio has elements of value and quality investing; (3) macro inefficiency is greater than micro inefficiency (consistent with Samuelson's Dictum) when there exists a strong common factor in security fundamentals or when the number of securities is large; (4) when information costs decline, the number of active managers increases, active fees decrease, market inefficiency decreases, especially macro inefficiency (counter to part of Samuelson's Dictum); (5) when the cost of passive investing decreases, market inefficiency increases, especially macro inefficiency, the number of active mangers decreases, and active fees drop by less than passive fees; and (6) market inefficiency is linked to the economic value of information and to entropy. These findings help explain a number of empirical findings and give rise to new tests as we discuss below.

To understand our results, let us briefly explain the framework. We introduce asset managers into the classic noisy-rational-expectations-equilibrium (REE) model of Grossman

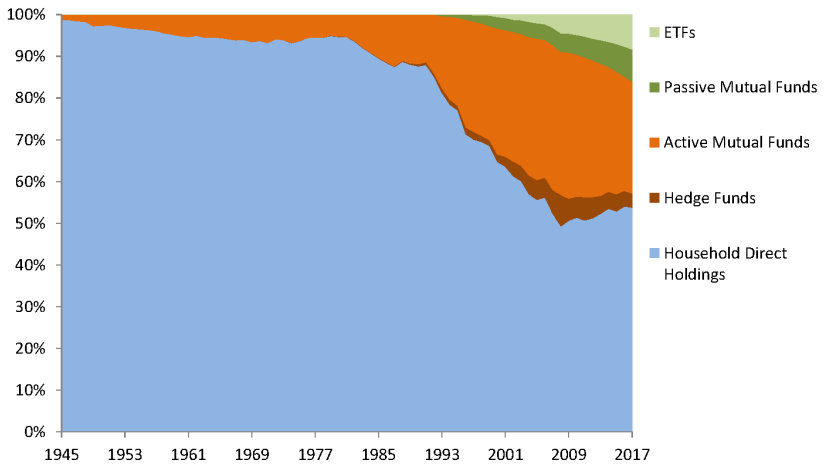

Figure 1: **Ownership of US Equities.** The figure shows the decline in direct holdings (blue) and the rise of delegated management, including active mutual funds and hedge funds (the two red areas), passive mutual funds and ETFs (the green areas), and pension plans and other equity owners (grey areas). The data is from the Federal Reserve's Flow of Funds Report, except the hedge fund data, which is from HFR, and the breakdown of mutual fund holdings into active versus passive, which is from Morningstar.

and Stiglitz (1980), following Gârleanu and Pedersen (2018). We extend the model to cover multiple securities (in the spirit of Admati (1985), who also considers multiple securities, but not asset managers) and costly asset management.[1] Having asset managers with multiple securities allows us to study portfolio choice and macro vs. micro efficiency and having costly active and passive management is essential to study the effects of changes in these costs as information technology evolves.

Investors must decide whether to invest on their own, allocate to a passive manager, or search for an active manager. Each of these alternatives is associated with a cost: self-directed trade has an individual-specific cost (time and brokerage fees), passive investing has a fee (equal to the marginal cost of passive management, in equilibrium), and active investing is associated with a search cost (the cost of finding and vetting a manager to

[1]Gârleanu and Pedersen (2018) features a single risky security and no passive managers, and assumes that self-directed trade is costless. See also García and Vanden (2009), who considers an alternative generalization of the noisy rational expectations framework with mutual funds trading a single risky asset. Influential papers focusing on other aspects of asset management include Berk and Green (2004), Petajisto (2009), Pastor and Stambaugh (2012), Vayanos and Woolley (2013), Berk and Binsbergen (2015), and Pastor et al. (2015).

ensure that she is informed) plus an active management fee. Active and passive managers determine which portfolios to choose and, in addition, active managers decide whether or not to acquire information. Market clearing requires that the total demand for securities equals the supply, which is noisy (e.g., because of share issuance, share repurchases, changes in the float, private holdings, etc.).

Passive managers seek to choose the best possible portfolio conditional on observed prices (but not conditional on the information that active managers acquire). One may wonder whether they should choose the "market portfolio" (the market-capitalization weighted portfolio of all assets), which is the standard benchmark in the Capital Asset Pricing Model (CAPM). While the market portfolio is the focal point of much of financial economics, it is *not* discussed in the context of REE models because supply noise renders it unobservable (likewise, in the real world no one knows the true market portfolio as emphasized by Roll (1977)); furthermore, much of the REE literature studies a single asset, precluding a meaningful analysis of portfolios. Bridging the REE literature and the CAPM, we show that passive investors choose the closest thing they can get to the market portfolio, namely the "expected market portfolio" based on the distribution of the supply and what can be learned from public signals, under certain conditions (e.g, i.i.d. shocks across securities). This behavior resembles what real-world passive investors do, namely choosing an index that is rebalanced based on public information. However, indices only hold a subset of all securities, typically large and mature firms with sufficient time since their initial public offering. In a similar spirit, we show that passive investors optimally over-weight securities with less supply uncertainty in the more general case when shocks are correlated across securities via a factor structure. Hence, our framework presents a first step toward a theory of optimal security indices.

Turning to active managers, we show that they use their information advantage in several ways. First, they naturally tend to overweight securities about which they have favorable information (quality investing). Second, they tend to overweight securities with positive supply shocks, since these securities fall in price (a form of value investing).

Active investors thus exploit market inefficiency across various assets, where inefficiency is defined (following Grossman and Stiglitz (1980)) as the uncertainty about the fundamental value conditional on only knowing the price relative to the uncertainty conditional on also knowing the private information. For example, the market inefficiency is zero (fully efficient market) if the uncertainty about the fundamental value is the same whether one learns from just the price or also the signal. Further, the more information advantage one enjoys from knowing the signal, the more inefficient the market.

An interesting question that can be addressed naturally in this framework is whether there are greater inefficiencies at the macro or at the micro level. Indeed, Samuelson famously hypothesized that macro inefficiency is greater than micro inefficiency, a notion known as "Samuelson's Dictum" (see quote and references in the beginning of Section 3 and the empirical evidence in Jung and Shiller (2005)). We show that Samuelson's Dictum holds when there exists a strong common factor in security fundamentals. More precisely, we show that the factor-mimicking portfolio is the most inefficient portfolio, while the least inefficient portfolios are long-short relative-value portfolios that eliminate factor risk. Hence, this makes precise what macro and micro efficiency means, and gives precise conditions under which Samuelson's Dictum applies (or does not apply). We further show that, due to diversification, when the number of securities is large, not only does Samuelson's Dictum always hold, but in fact the combined inefficiency of all micro portfolios becomes negligible — all the inefficiency is in the pricing of systematic factors. This result is related to the Arbitrage Pricing Theory (APT) of Ross (1976). While the APT states that risk premia are driven by systematic factors when the number of assets is large, we show that inefficiencies are also driven by these factors.[2]

[2]Active managers in our model make an all-or-nothing information choice, whereas Veldkamp (2011), Van Nieuwerburgh and Veldkamp (2010), Kacperczyk et al. (2016), Glasserman and Mamaysky (2018), and others study agents' choice of information, which can affect macro vs. micro efficiency as emphasized by the latter paper. Our assumption captures, for example, the case in which active investors decide whether or not to set up an IT system that captures the main databases, whereas the above papers capture the idea that different investors may focus on different subsets of the available information. See also Breugem and Buss (2018), who consider the effect of benchmarking considerations on information acquisition and efficiency with multiple assets, and Kacperczyk et al. (2018), who consider the effect of large investors' market power on market efficiency. We complement the literature with regard to macro vs. micro efficiency by providing a general definition of Samuelson's Dictum, by showing how it arises with many assets (i.e., as the number of

Samuelson also hypothesized that efficiency, especially micro efficiency, has improved over time (see quote and reference in Section 4). Such an improvement in efficiency may be driven by a reduction in information costs as information technology has improved. We show that reduced information costs indeed lower inefficiencies, but they actually mostly lower macro inefficiency (counter to that part of Samuelson's hypothesis). Lower information costs also increase active management (relative to self-directed investment and passive management), consistent with the development in the 1980s and 1990s.

Another trend over the past decades is the decline in the cost of passive management. We show that such a decline should lead to a rise in passive management (at the expense of self-directed investment and active management), consistent with the development in the 2000s. Further, reduced cost of passive management leads to an increase in market inefficiency, especially macro inefficiency, leading to stronger performance of active managers so that their fees decrease by less than passive fees. These predictions are consistent with the empirical findings by Cremers et al. (2016). Indeed, Cremers et al. (2016) find that lower-cost index funds leads to a larger share of passive investment, higher average alpha for active managers, and lower fees for active, but an increased fee spread between active and passive managers.

Finally, we show that market inefficiency is linked to the economic value of information and to relative entropy (and the Kullback-Leibler divergence), which provides a new way to estimate efficiency.

In summary, we complement the literature by studying the optimal passive and active portfolios and the nature of optimal security indices, giving general conditions for Samuelson's Dictum, and studying what happens when information costs and asset management-costs change. Our model provides a financial economics framework that links the CAPM, APT, and REE in a way that helps explain recent trends in financial markets and in the financial services sector. Finally, we quantify the model's implications via a calibration.[3]

assets goes to infinity), and by showing the importance of all systematic factors (not just the market, assumed exogenously by Glasserman and Mamaysky (2018)), thus linking to APT. More broadly, we complement the literature by capturing investors' costs of active, passive, and direct investments, by studying the effects of changes in the asset-management costs, and by relating efficiency to entropy.

[3]Stambaugh (2014) also considers trends in the investment management sector based on a different framework where the key driving force is a reduction in the amount of noise trading. As noise trading

1 Model and Equilibrium

This section lays out our noisy rational expectations equilibrium (REE) model and shows how to solve it.

1.1 REE Model with Multiple Assets and Asset Managers

We model a two-period economy featuring a risk-free security and n risky assets. The return of the risk-free security is normalized to zero while the vector risky asset prices p is determined endogenously. The risky assets deliver final payoffs given by the vector v, which is normally distributed with mean $\bar{v}$ and variance-covariance matrix Σ_v, which we write as $v \sim \mathcal{N}(\bar{v}, \Sigma_v)$.

Agents can acquire various signals about all the assets at a cost k. We collect all the signals in a vector of dimension n that we denote $s = v + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$ is the noise in the signal.[4]

The supply of the risky assets is given by $q \sim \mathcal{N}(\bar{q}, \Sigma_q)$ and the shocks to q, ε, and v are independent. The supply is noisy for several reasons (e.g., Pedersen, 2018): New firms are listed in initial public offerings, existing firms issue new shares in seasoned equity offerings, firms repurchase shares (sometimes by buying shares in the market without telling investors), and the number of floating shares changes when control groups buy or sell shares.[5] Further, the de facto supply of publicly traded shares also implicitly changes when correlations vary between public shares and investors' endowment (e.g., human capital, natural resources, or private equity holdings, where private firms may also issue or repurchase shares). For these

declines in his model, the allocation to, and the performance of, active managers both decline. Hence, this model cannot explain the finding of Cremers et al. (2016) discussed above, namely that the size and performance of active management move in opposite directions (but the model can explain a number of other phenomena).

[4] If we start with a signal $\hat{s}$ of any other dimension $\hat{n}$, then we can focus on the conditional mean $u := \mathrm{E}(v|\hat{s})$, which is of dimension n and can be translated into a signal s as modeled above. For example, if $\hat{n} \geq n$, we have $s := \bar{v} + \Sigma_v \Sigma_u^{-1}(u - \bar{v})$, where $\Sigma_u = \mathrm{var}(u)$.

[5] Many indices use a float adjustment. E.g., S&P Float Adjustment Methodology 2017 states "the share counts used in calculating the indices reflect only those shares available to investors rather than all of a company's outstanding shares. Float adjustment excludes shares that are closely held by control groups, other publicly traded companies or government agencies."

and other reasons, no one knows the true market portfolio, the underpinning of the "Roll critique" (Roll, 1977).

The economy has $\bar{M}$ active asset-management firms and a representative passive manager. The passive manager seeks to deliver the best possible portfolio that can be achieved without acquiring the signal s. The passive manager faces a marginal cost per investor of k_p and, since passive investing is assumed to be a competitive industry, this manager charges a fee $f_p = k_p$ (where the subscript p naturally stands for "passive").

Active managers face a marginal cost of k_a (subscript a for "active") and, in addition, they must decide whether to incur the fixed cost k associated with acquiring the signal s. Specifically, M active managers endogenously decide to pay the cost k to become informed while the remaining $\bar{M} - M$ managers seek to collect active asset management fees f_a even though they invest without information (e.g., these managers are using "closet indexing").

The economy has $\bar{S}$ optimizing investors with initial wealth W and constant absolute risk aversion (CARA) coefficient γ. These investors either search for an active manager, allocate to a passive manager, or invest directly in the financial market. Specifically, S_a investors choose to search for an active manager, S_p investors choose passive, and the remaining $\bar{S} - S_a - S_p$ are self-directed. If an investor l makes an uninformed investment (i.e., without using the signal s) directly in the financial market, then he incurs an investor-specific cost d_l associated with brokerage fees and time used on portfolio construction. If the investor makes an informed investment, the cost is $d_l + k$, but we show that this behavior is dominated by using an active manager. If the investor uses a passive manager, he incurs the passive management fee f_p (as discussed above). Finally, investors can allocate to an informed active manager by paying a search cost c to be sure to find an informed manager and, in addition, pay an asset management fee f_a determined via Nash bargaining.

The search cost $c(M, S_a)$ is a smooth function of the number of searching investors S_a and the number of informed managers M. For a number of our results, it is helpful that the search cost satisfies the regularity conditions $\frac{\partial c}{\partial M} \leq 0$ and $\frac{\partial c}{\partial S_a} \geq 0$, namely that it is easier to find an appropriate manager if there are more managers and harder when there are more

agents searching for a manager. We maintain this assumption throughout.[6]

Finally, the economy has $N \geq 0$ "noise allocators" who invest with random active asset managers, e.g., due to issues related to trust as modeled by Gennaioli et al. (2015). These investors are not important for the model, but they allow the possibility that uninformed active managers have some clients (which could alternatively be achieved by having a less than perfect search technology) and, moreover, the existence of trust-based investors may be empirically realistic (as also discussed in our calibration in Section 6).

The total number of active investors is therefore the sum of the searching investors and the noise allocators, $S_a + N$. Of these active investors, the total number of investors I with informed managers equals the searching investors S_a plus a fraction of the noise allocators, $I \equiv S_a + N\frac{M}{\bar{M}}$. In particular, as a consequence of their random manager choice, a proportion $\frac{M}{\bar{M}}$ of the noise allocators end as invested with an informed active manager. The remaining investors, $U \equiv \bar{S} + N - I$, remain uninformed.

In summary, we look for an equilibrium defined as follows. An equilibrium consists of a vector of asset prices p, an active asset management fee f_a, a number of informed active managers M, and numbers of optimizing investors who allocate to active S_a or passive managers S_p such that: (i) the supply of shares equals the demand; (ii) fees are set via Nash bargaining; (iii) each active manager decides optimally whether to be informed; and (iv) each investor decides optimally whether to use an active manager, use a passive manager, or be self-directed. Finding an equilibrium therefore entails solving $n + 4$ equations with as many unknowns, namely n market clearing conditions (one for each risky asset), one manager indifference condition, one fee-determination equation, and two investor conditions. We show below how to solve the model in a straightforward way.[7]

[6] An analytically convenient function that satisfies these requirements is $c(M, S_a) = \bar{c}\left(\frac{S_a}{M}\right)^{\alpha}$ for constants $\bar{c} > 0$ and $\alpha > 0$.

[7] In general, an equilibrium with non-zero S_a, S_p, and M need not be unique, but we concentrate throughout on the equilibrium featuring the largest value of I and assume throughout parameters for which the largest equilibrium is interior, in that the numbers of each of the three types of investors are strictly positive. In a related set-up, Gârleanu and Pedersen (2018) discusses in detail equilibrium determination and multiplicity.

1.2 Efficiency of Assets, Portfolios, and the Market

An important building block of our analysis is the notion of price efficiency. To define this concept, we build on the logic of Grossman and Stiglitz (1980), which considers the inefficiency of a single asset.

We wish to define the inefficiency of any set of linearly independent portfolios $\{\zeta_1, ..., \zeta_l\} \subset \mathbb{R}^n$, where the number of portfolios can be anywhere from $l = 1$, i.e., a single asset, to $l = n$, that is, the entire market. We collect the portfolio weights in a matrix $\zeta \in \mathbb{R}^{n \times l}$ and define their joint inefficiency as follows.

$$\eta^{\zeta} = \frac{1}{2} \log \left(\frac{\det(\text{var}\left(\zeta^{\top} v | \mathcal{F}_u\right))}{\det(\text{var}\left(\zeta^{\top} v | \mathcal{F}_i\right))} \right), \tag{1}$$

where $\mathcal{F}_i = \mathcal{F}(p, s)$ is the informed information set, consisting of both the price and the signal, and $\mathcal{F}_u = \mathcal{F}(p)$ is the uninformed information set, consisting only of the price.

In words, this definition means that a set of portfolios is considered more inefficient if the uninformed has a larger uncertainty relative to the informed about the fundamental values of these portfolios. For example, the inefficiency of a single asset, say asset 1, is computed by considering the portfolio $\zeta = (1, 0, ..., 0)^{\top}$, which yields

$$\eta^{\text{asset 1}} = \frac{1}{2} \log \left(\frac{\text{var}\left(v_1 | \mathcal{F}_u\right)}{\text{var}\left(v_1 | \mathcal{F}_i\right)} \right) = \log \left(\frac{\text{var}\left(v_1 | \mathcal{F}_u\right)^{1/2}}{\text{var}\left(v_1 | \mathcal{F}_i\right)^{1/2}} \right). \tag{2}$$

This expression is equivalent to that of Grossman and Stiglitz (1980). The expression makes the link between our notion of inefficiency and the amount of information gleaned from signals, respectively only prices, particularly easily to see.

The overall market inefficiency plays a special role in the equilibrium of the model. The overall market inefficiency is naturally the inefficiency of the set of all assets. Hence, we consider the largest possible matrix of portfolios, $\zeta = I_n$, namely the identity matrix in

$\mathbb{R}^{n\times n}$.[8] We denote the overall market inefficiency simply by η:

$$\eta = \eta^{\text{overall market}} = \eta^{I_n} = \frac{1}{2}\log\left(\frac{\det(\text{var}\,(v|\mathcal{F}_u))}{\det(\text{var}\,(v|\mathcal{F}_i))}\right). \tag{3}$$

This definition of overall market inefficiency is the natural extension of the one-asset definition of Grossman and Stiglitz (1980), since it retains the tight link between market inefficiency and investors' utility of information as discussed further below and formalized in Proposition 8.[9] Another natural property of our notion of market inefficiency is that it is linked to entropy, which is also stated in Proposition 8.

When we analyze macro vs. micro efficiency (in Section 3), we also study the efficiency of individual securities, portfolios, and collections of portfolios; the general definition of efficiency will be very useful.

1.3 Solution: Deriving the Equilibrium

To solve for an equilibrium, one proceeds backwards in time. The first step, therefore, consists of solving for an equilibrium in the asset market, taking as given the masses of investors conditioning on $\mathcal{F}_i$, respectively on $\mathcal{F}_u$. This is done in a standard Grossman-Stiglitz step. We conjecture and verify that prices p are linear in the information s about securities as well as the supply q:

$$p = \theta_0 + \theta_s\left((s-\bar{v}) - \theta_q(q-\bar{q})\right). \tag{4}$$

The resulting optimal demands are linear, as well. For an investor of type $j \in \{i,u\}$, where i means that the investor invests through an informed manager and u means that he invests uninformed (passive manager or self-directed), the optimal demand is the portfolio x_j that maximizes the investor's expected utility given the information used.[10] The resulting

[8] The same outcome for the overall market inefficiency obtains for any matrix $\zeta \in \mathbb{R}^{n\times n}$ of full rank.

[9] The link between the value of information and the ratio of determinants of the conditional variances was first derived in Admati and Pfleiderer (1987).

[10] When an investor has searched for a manager, confirmed that the manager is informed, and paid the fee, then the manager invests in the investor's best interest. This lack of agency problems means that there

certainty equivalent utility is

$$-\frac{1}{\gamma}\log\left(\mathrm{E}\left[\max_{x_j}\mathrm{E}\left(e^{-\gamma\left(W+x_j^\top(v-p)\right)}\,|\mathcal{F}_j\right)\right]\right) =: W+u_j, \tag{5}$$

where $\mathcal{F}_j$ is the information set of investor j (as defined above). The above equality defines the certainty equivalent utility of being informed, u_i, respectively uninformed, u_u, as well as the corresponding optimal portfolios x_i and x_u. With these portfolio choices by informed and uninformed investors, market clearing requires that the supply of shares q equals the total demand,

$$q = Ix_i + Ux_u, \tag{6}$$

where the number of informed investors (I) and the number of uninformed ones (U) are defined in Section 1.1.

Despite the presence of many assets, the overall asset-market equilibrium is summarized by a single number, namely the overall inefficiency η defined in equation (3). Indeed, the market inefficiency captures investors' utility of information:

$$\gamma(u_i - u_u) = \eta. \tag{7}$$

Thus, when the overall market is more inefficient, investors have a greater utility gain from being informed relative to being uninformed (as we show in Proposition 8).

The second step consists of determining the active-management fee, taking into account the utility consequences of investing actively with an informed manager, respectively passively. The active asset management fee is set through Nash bargaining, meaning that the fee maximizes the product of the manager's and investor's gains from trade. The investor's gain from his investment is his certainty equivalent utility if he invests $(W - c - f_a + u_i)$ over and above his outside option of going to a passive manager $(W - c - f_p + u_u)$, where c

is no difference between investing in a fund and sale of information as in Admati and Pfleiderer (1990). For a recent model of agency issues in asset management, see Buffa et al. (2014).

appears in both terms because it is a cost that is already sunk.[11] The manager's gain from accepting the investor is her fee revenue f_a less his marginal cost k_a. Hence, the equilibrium fee is:

$$\begin{aligned} f_a &= \arg\max_f \left(W - c - f + u_i - \left(W - c - f_p + u_u\right)\right)\left(f - k_a\right) \\ &= \arg\max_f \left(u_i - u_u + f_p - f\right)\left(f - k_a\right) = \frac{u_i - u_u + f_p + k_a}{2} \\ &= \frac{k_a + k_p}{2} + \frac{\eta}{2\gamma}, \end{aligned} \tag{8}$$

where we use (7) and the equality of the passive management fee and the marginal cost $f_p = k_p$.

We see that the equilibrium active asset management fee f_a equals the average marginal cost of active and passive asset management plus a term that increases in the market inefficiency η. Intuitively, active managers can add more value in a more inefficient market, and hence charge larger fees.

The third step involves the ex-ante considerations of managers and investors. Let's start with the managers. A manager that remains uninformed only attracts a share of the noise allocators, with an expected value of $N/\bar{M}$. If she acquires the signal instead, and becomes an informed active manager, then she expects S_a/M additional investors, giving rise to an extra income, net of marginal costs, of $(f_a - k_a)S_a/M$. Hence, the active manager's indifference condition for paying the information cost k is

$$\frac{\eta}{2\gamma} + \frac{k_p - k_a}{2} = \frac{M}{S_a}k. \tag{9}$$

As for the investors, their optimal allocations are determined as follows. An investor l with a low cost direct investment $d_l < f_p$ optimally invest directly in the financial market. Investors with higher costs of direct investment $d_l \geq f_p$ are indifferent between active and passive management in an interior equilibrium. The indifference condition for these investors

[11] The investor's outside option can also be seen as searching again for another active manager, which yields the same result in an interior equilibrium.

equalizes the certainty equivalent utility of passive management ($W + u_u - f_p$) with that of active management ($W + u_i - c - f_a$),

$$u_i - u_u = f_a - f_p + c \tag{10}$$

which can be rewritten using (7) and (8) as:

$$\frac{\eta}{\gamma} = k_a - k_p + 2\,c. \tag{11}$$

The equilibrium condition (10)–(11) is intuitive. It says that the benefit of informed investing (the left-hand side) must equal the net cost of being informed (the right-hand side). The benefit equals the gain from exploiting market inefficiency, η, which we divide by γ to measure in terms of certainty equivalent dollars. The net cost of active investing is the active fee plus the search cost minus the passive fee. This net cost can be reduced to the difference in marginal costs, $k_a - k_p$, plus twice the search cost (twice because active investors must both pay the search cost and the active fee, and the latter equals the search cost plus the marginal cost, $f_a = c + k_a$, in equilibrium).

Figure 2 summarizes the procedure for finding an interior equilibrium by equalizing the cost and benefit of active investing. Starting with the benefit, we first compute the left-hand side of (11) for each number of informed investors, I, generating the solid line in the figure. To do this, we solve for the asset prices (4) using the optimal demand (5) and asset-market clearing (6). This asset-market equilibrium yields the inefficiency η, given by (3), for each I.

Turning to the cost of active investing, we compute the right-hand side of (11) for each number of informed investors, I, generating the dashed line in Figure 2. The search cost $c\,(S_a, M)$ depends on the number of searching investors S_a and the number of informed managers M they are looking for, so we first derive (S_a, M) for each I. In particular, we see that (S_a, M) can be derived uniquely by combining $I = S_a + N\frac{M}{M}$ with the manager indifference condition (9). We can therefore think of this cost as the actual cost that the investor would face, for any I, as long as the managers are in equilibrium for that I.

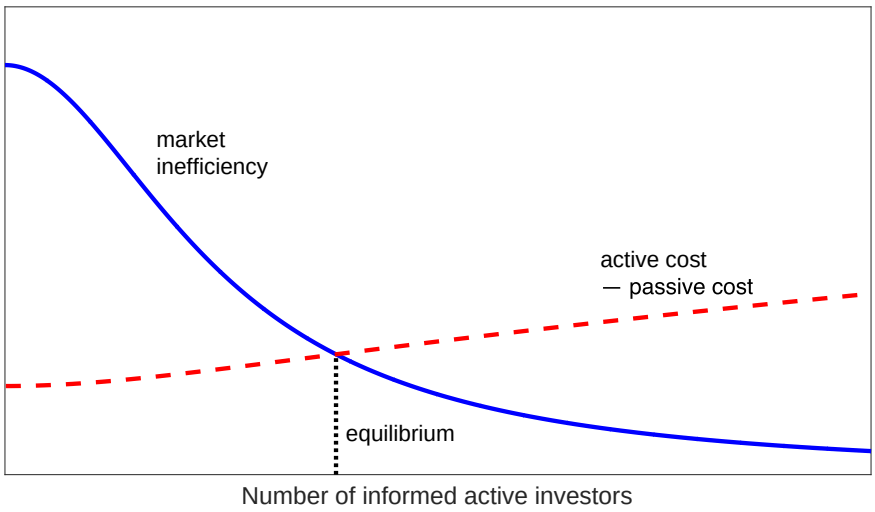

Figure 2: **Equilibrium Market Inefficiency.** The figure shows how an equilibrium is found by equalizing the cost and benefit of active investing. The benefit of active investing (the solid line) stems from the ability to exploit market inefficiency; expressed as certainty equivalent wealth, it equals η/γ. The net cost of active investing (the dashed line) equals the active fee plus the search cost, minus the cost of passive investing.

Finally, as seen in Figure 2, the equilibrium is found as the intersection of the solid and dashed lines. Intuitively, having more investors with informed managers reduces market inefficiency, diminishing the benefit of being informed. Equalizing this benefit with the cost of active investing yields the equilibrium. This intuitive figure makes it easy to investigate the implications of varying parameters; we'll find it useful, for instance, when we consider the effect of changing passive costs in Section 4.

1.4 Statistical Assumptions

Before we discuss our main results, we note that certain properties of the equilibrium (such as investors' portfolios and the validity of the Samuelson's Dictum) simplify when we impose further statistical structure on the shocks. We mainly rely on Assumption 1, namely that the covariance of shocks can be represented by a factor model.

Assumption 1 *Fundamentals have a factor structure:*

$$v = \bar{v} + \beta F_v + w_v \tag{12}$$

$$\varepsilon = \beta F_\varepsilon + w_\varepsilon \tag{13}$$

$$q = \bar{q} + \beta F_q + w_q, \tag{14}$$

where $\bar{v}, \bar{q} \in \mathbb{R}^n$ are the average fundamental values, respectively supplies, $\beta \in \mathbb{R}^n$ is a vector of factor loadings normalized (without loss of generality) such that $\beta^\top \beta = n$, the common factors F_v, F_ε, and F_q are one-dimensional random variables with zero means and variances $\sigma^2_{F_v}$, $\sigma^2_{F_\varepsilon}$, and $\sigma^2_{F_q}$, respectively, and the idiosyncratic shocks w_v, w_ε, and $w_q \in \mathbb{R}^n$ are i.i.d. across assets with variances $\sigma^2_{w_v}$, $\sigma^2_{w_\varepsilon}$, respectively $\sigma^2_{w_q}$ for each asset.

We will refer to the portfolio proportional to β as the *"factor portfolio,"* since this portfolio is maximally correlated with the common shocks. It is natural to think of this factor as the average market portfolio — that is, under Assumption 1 it is natural to also assume that the average supply $\bar{q}$ is proportional to β and we will occasionally make this additional assumption.

We also make use of the following assumption, which captures the idea that may underlie Samuelson's hypothesis, namely that the common factor-component of the risk is especially important for future security prices.

Assumption 1′ *Assumption 1 holds and the common factor of v is non-zero, $\sigma^2_{F_v} > 0$, and at least as important as that of ε, i.e., $\sigma^2_{F_v}/\sigma^2_{w_v} \geq \sigma^2_{F_\varepsilon}/\sigma^2_{w_\varepsilon}$.*

Finally, we also consider the following assumption.

Assumption 2 *There exist scalars z_ε and z_q such that $\Sigma_\varepsilon = z_\varepsilon \Sigma_v$ and $\Sigma_q^{-1} = z_q \Sigma_v$.*

The first part of Assumption 2 simply says that fundamentals and signal noise have the same risk structure (which can also hold under Assumption 1). The second part, which is

more unusual, says that the inverse of the variance-covariance matrix of the supply noise also shares this structure.[12]

Assumptions 1 and 2 are both satisfied if all shocks are i.i.d. across assets, but otherwise they are different. We focus on Assumption 1, as it is the more standard and more realistic assumption. Assumption 2 is to be thought of as a generalization of the i.i.d.-shock case. In particular, the results that require narrowing Assumption 1 down to the case of i.i.d. shocks also hold under Assumption 2, and we therefore state them in this greater generality.

2 Optimal Passive and Active Portfolios

We wish to understand how active and passive investors construct their portfolios. Specifically, we are interested in the optimal informed portfolio x_i and uninformed portfolio x_u, defined in Equation (5). Given that real-world uninformed investors tend to hold indices, the optimal uninformed portfolio provides a foundation for the economics of indices.

A standard benchmark portfolio in financial economics is the "market portfolio," that is, the portfolio of all assets (cf. the CAPM). However, as emphasized by Roll (1977), the market portfolio is not known in the real world. Likewise, uninformed investors do not know the market portfolio q in our noisy REE economy.

While the market portfolio is central in the CAPM, it has not played a role in the REE literature (because this portfolio is public knowledge and because this literature is focused on a single asset). Nevertheless, we can bridge these literatures by introducing the concepts of the *"conditional expected market portfolio,"* $E(q|p)$, and the *"average market portfolio,"* $\bar{q}$. The conditional expected market portfolio is the uninformed investors' best estimate of the true market portfolio, q, based on public information. We first show how the optimal

[12]To understand Assumption, consider what happens if any security j undergoes a two-for-onw stock split, meaning that all shareholders receive two new shares for each old share. In this case, the number of shares outstanding doubles and the value of each share drops by half. This means that, if Assumption 2 was satisfied before the stock split, then it remains satisfied after the stock split. Indeed, the split means that the volatility of the value of shares drops by half, the volatility of the information noise drops by the same ratio, and the volatility of the supply noise doubles. A less natural implication of Assumption 2 is that securities with more correlated fundamentals have less correlated supply shocks (except in the special case, which overlaps with Assumption 1, when all securities are i.i.d.).

uninformed (or passive) portfolio is linked to these notions of the market portfolio.

Proposition 1 (Optimal passive portfolio: the economics of an index)

1. *The optimal uninformed portfolio, x_u, is related to the conditional expected market portfolio, $E(q|p)$, in the following way:*

   a. *The portfolio x_u is proportional to $E(q|p)$, in that there exists $A \in \mathbb{R}^{n\times n}$ such that*

   $$x_u = A\, E(q|p). \tag{15}$$

   b. *Under Assumption 2, $A \in \mathbb{R}_+$ is a scalar.*

   c. *Under Assumption 1, A is positive definite and scalars $A_0 > 0$ and A_1 exist such that*

   $$x_u = A_0\, E(q|p) - A_1\, \beta^\top E(q|p)\beta. \tag{16}$$

   Further, $A_1 > 0$ if Assumption 1′ holds.

2. *The average passive portfolio, $E(x_u)$, is related to the average market portfolio, $\bar{q}$, as follows.*

   a. *Under Assumption 2 or if Assumption 1 holds and β and $\bar{q}$ are collinear, then $\bar{A} \in \mathbb{R}_+$ exists such that*

   $$E(x_u) = \bar{A}\, \bar{q}. \tag{17}$$

   b. *Normalize β so that $\beta^\top \bar{q} \geq 0$. Then, under Assumption 1′, if two assets have the same weight in the average market portfolio, $\bar{q}_i = \bar{q}_j$, but different risk, $\beta_i > \beta_j$, then the safer asset will have a larger weight in the average passive portfolio, $E(x_{u,i}) \leq E(x_{u,j})$.*

c. *If v, ϵ, and q are i.i.d. across assets, except that the supply uncertainty $(\Sigma_q)_{jj}$ varies across assets indexed by j, then passive investors hold larger average positions $E(x_{u,j})$ in assets with lower supply uncertainty.*

It is interesting to compare the conclusions of this proposition to the actual portfolios of real-world passive investors. Passive investors typically hold a combination of (global) stock and bond index funds or ETFs. These indices are typically weighted by the total market-capitalization of each constituent security's floating shares outstanding. The indices are regularly re-weighted to account for changes in the index constitution, share repurchases, new issuance, and changes in the insider holdings (i.e., shares not part of the float). Hence, consistent with part 1.a of Proposition 1, passive indices seek to use public information to proxy for the expected market portfolio as an input to the portfolio construction. Further, part 1.b states that, under certain conditions, the optimal passive portfolio is literally just that, namely the conditional expected market portfolio.

Part 1.c shows that, under the more realistic Assumption 1, the optimal passive portfolio is similar to the conditional expected market portfolio, but tilted away from risky securities. This tilt arises because uninformed investors face an extra risk (relative to informed investors) due to supply uncertainty.

Part 2.a shows further that, when the average market portfolio is proportional to the factor portfolio, $\beta \sim \bar{q}$, then the average passive portfolio is also proportional to the average market portfolio.

Part 2.b shows that, in the more general situation in which β and $\bar{q}$ are not proportional, the optimal passive portfolio tends to downweight risky securities. Part 2.c conveys a similar intuition (under different assumptions), again showing that passive investors optimally downweight securities with more supply uncertainty. This intuition may help explain why real-world passive indices only include a subset of listed securities, which may be interpreted as a binary version the result of parts 2.b and 2.c of the proposition. While the proposition states that passive investors will hold more of securities with less supply uncertainty, real-world passive investors may exclusively hold these securities. Indeed, indices typically exclude

securities with too low price, too low market capitalization, too low liquidity, too recent IPO, or involved in certain corporate actions.[13]

Finally, we note that portfolio holdings scale with the degree of risk aversion. The results therefore only depend on the actual risk aversion of any given investor through a (positive) constant of proportionality.

We next turn to the portfolio holdings of the informed investors.

Proposition 2 (Optimal active portfolio: value and quality) *Under Assumption 1 or Assumption 2, an informed investor's position in any asset j is more sensitive than that of an uninformed agent to*

(a) supply shocks for asset j, $\frac{\partial E[x_{i,j}|q]}{\partial q_j} > \frac{\partial E[x_{u,j}|q]}{\partial q_j}$ *(value investing);*

(b) the signal s_j about asset j, $\frac{\partial E[x_{i,j}|s]}{\partial s_j} > 0 > \frac{\partial E[x_{u,j}|s]}{\partial s_j}$ *(quality investing).*

The first part of the proposition states that informed investors buy more when the supply increases. For example, when there is an initial public offering, informed investors likely buy a disproportional fraction of the shares during book-building process. Likewise, if the supply of an existing company increases (for a given value of the signal), this extra supply will tend to lower the price, creating buying opportunity for informed investors (who realize that the price drop is not due to bad information). Buying securities at depressed prices can be viewed as a form of "value investing."

The second part of the proposition states that, when the signal for a given security improves, informed investors tend to increase their position in this security while uninformed investors tend to lower their position. Clearly, when informed investors receive favorable information about a security, they are more inclined to buy it. This extra demand tends to increase the price, leading the uninformed to reduce their position (markets must always clear) since uninformed investors cannot know whether the price increase due to favorable information or a drop in supply. Buying securities with strong fundamentals, even if their price has increased, is called "quality investing."

[13]See, e.g., and "Russell U.S. Equity Indexes 2017" and "S&P U.S. Indices Methodology 2017."

The idea that informed investors should focus on value and quality goes back at least to Graham and Dodd (1934) and, following this advice, investors such as Warren Buffett have pursued these strategies (Frazzini et al. (2018)). Value and quality investment strategies have indeed been profitable on average across global markets (see, e.g., Asness et al. (2013), Asness et al. (2018), Fama and French (2017)).

3 Samuelson's Dictum: Macro vs. Micro Efficiency

We have seen that the overall level of market efficiency is linked to the level of active asset management fees in (8). It is interesting to further study the relative price efficiency of different securities and portfolios. In this connection, Paul Samuelson famously conjectured that markets would have greater micro efficiency than macro efficiency:[14]

> *"Modern markets show considerable micro efficiency (for the reason that the minority who spot aberrations from micro efficiency can make money from those occurrences and, in doing so, they tend to wipe out any persistent inefficiencies). In no contradiction to the previous sentence, I had hypothesized considerable macro inefficiency, in the sense of long waves in the time series of aggregate indexes of security prices below and above various definitions of fundamental values."*

Our framework is an ideal setting to make Samuelson's intuition precise. Indeed, we have multiple securities (so we can discuss micro vs. macro) and a precise measure of efficiency for any asset or portfolio given by (1).

Inspired by Samuelson's Dictum, we are interested in which portfolio ζ has the maximum inefficiency:

$$\max_{\zeta\in\mathbb{R}^n} \eta^\zeta = \max_{\zeta\in\mathbb{R}^n} \frac{1}{2}\log\left(\frac{\zeta^\top \mathrm{var}(v|p)\zeta}{\zeta^\top \mathrm{var}(v|s)\zeta}\right).$$

[14]This quote is from a private letter from Samuelson to John Campbell and Robert Shiller, as discussed by Shiller (2001). Other references to the notion of macro vs. micro efficiency appear in, e.g., Samuelson (1998).

We wish to determine whether the solution, say ζ^*, is micro or macro in nature. Similarly, we are interested in which portfolio has the minimum inefficiency, but since the analysis is analogous, we focus here on the maximum and state the general result in the proposition below.

To solve this problem, we let $G = \text{var}(v|s)^{-1/2}\text{var}(v|p)\text{var}(v|s)^{-1/2}$, which is essentially the matrix of the informed investor's information advantage (in terms of her reduction in uncertainty). We denote its eigenvalues by $g_1 \geq g_2 \geq ... \geq g_n > 0$ and the corresponding eigenvectors by w_1 through w_n. Using this matrix, we see that the maximum portfolio inefficiency is:

$$\max_{\zeta \in \mathbb{R}^n} \eta^\zeta = \max_{z \in \mathbb{R}^n} \frac{1}{2} \log \left(\frac{z^\top G z}{z^\top z} \right) = \frac{1}{2} \log g_1,$$

where we have used the substitution $z = \text{var}\,(v|s)^{1/2}\,\zeta$. The most inefficient portfolio is the eigenvector w_1 corresponding to the largest eigenvalue, translated back into portfolio coordinates (i.e., reversing the substitution), $\hat{w}_1 := \text{var}\,(v|s)^{-1/2}\,w_1$.

We have almost answered Samuelson's question namely whether the most inefficient portfolio, $\zeta^* = \hat{w}_1$, is macro or micro in nature. All that is left is to determine the portfolio $\hat{w}_1$ by finding the primary eigenvector of G.

Before we state the answer, we note that this analysis also sheds new light on the meaning of "overall market inefficiency" in an economy with multiple assets. Specifically, we can express the overall market inefficiency in terms of the eigenvalues (g_j):

$$\eta = \frac{1}{2} \log \left(\frac{\det(\text{var}(v|p))}{\det(\text{var}(v|s))} \right) = \frac{1}{2} \log \left(\det(G) \right) = \frac{1}{2} \sum_{j=1}^{n} \log g_j = \sum_{j=1}^{n} \eta^j, \tag{18}$$

where η^j is the inefficiency of portfolio $\hat{w}_j$, defined analogously to $\hat{w}_1$, as the "rotated" portfolio versions of eigenvector w_j: $\hat{w}_j = \text{var}\,(v|s)^{-1/2}\,w_j$. We see that the overall market inefficiency is the sum of portfolio inefficiencies for the set $(\hat{w}_j)$ of independent[15] portfolios

[15]While the original eigenvectors (w_j) are orthogonal in the Euclidean norm, $w_j^\top w_k = 0$, the corresponding portfolios $(\hat{w}_j)$ are orthogonal in the economically more interesting sense that their payoffs are conditionally uncorrelated, $\text{cov}(\hat{w}_j^\top v^\top, \hat{w}_k^\top v|s) = \hat{w}_j^\top \text{var}\,(v|s)\,\hat{w}_k = 0$. (We also note that, under Assumption 1, the

that spans the space of all portfolios.

To understand the economics of equation (18), note first that, if an informed investor can choose her "loading" on the most inefficient portfolio $\hat{w}_1$ (i.e., go long or short and scale the position up or down as she wishes), then she will get an expected utility benefit (relative to an uninformed investor) of $\frac{1}{2}\log g_1$.[16] Second, if the informed investor can also choose her loading on the second most inefficient portfolio, $\hat{w}_2$, then she will get an additional utility benefit of $\frac{1}{2}\log g_2$, and this utility is additive because of the independence of these portfolio returns and the CARA utility. Third, if the investor can choose her portfolio freely, which we can think of as the sum of loadings on all the "basis" portfolios ($\hat{w}_j$), then her expected utility (again, relative to that of an uninformed) is $\eta = \frac{1}{2}\sum_i \log g_i$.

We are ready to state our general result on macro vs. micro efficiency.

Proposition 3 (Macro vs. Micro Efficiency)

(a) Under Assumption 2, all portfolios are equally inefficient, i.e., η^ζ is the same for all portfolios $\zeta \in \mathbb{R}^n$.

(b) **(Samuelson's Dictum)** *Under Assumption 1′, the most inefficient portfolio is the factor portfolio,*

$$\max_{\zeta\in\mathbb{R}^n} \eta^\zeta = \eta^\beta$$

and the least inefficient portfolios are those that eliminate factor risk, i.e.,

$$\min_{\zeta\in\mathbb{R}^n} \eta^\zeta = \eta^z$$

for any z with $z^\top\beta = 0$.

(c) There exist parameters for which the opposite conclusion of part (b) holds.

rotation only scales the eigenvectors w_1 through w_n and the corresponding portfolios remain orthogonal in the Euclidian norm.)

[16] We also note that, in a world in which an informed investor was allowed only to choose her loading on a single portfolio fixed ex ante, then the information would be most valuable if this portfolio were the most inefficient one, $\hat{w}_1$, since this would result in the highest expected utility differential, $\frac{1}{2}\log g_1$.

(d) For all parameters satisfying Assumption 1, one of the above three conclusions applies, that is, either all portfolios are equally efficient, the factor portfolio is the most efficient, or the factor portfolio is the least efficient.

We see that, naturally, Samuelson's Dictum does not hold for all parameters. Part (a) of the proposition states that, under certain conditions, all portfolios are equally efficient. In other words, in this case, the market is equally efficient in a macro sense and a micro sense, counter to Samuelson's dictum. This conclusion applies, for example, when all securities are independent (in terms of the fundamental values, signals, and supply shocks).

Part (b) of the proposition gives a precise meaning to Samuelson's Dictum (in the context of our rational, information-based model). Specifically, we see that when we maximize and minimize inefficiency, the solutions turn out to be "macro" and "micro" portfolios, as conjectured by Samuelson. Further, the proposition provides conditions under which Samuelson's Dictum applies. The sufficient condition appears empirically plausible since it states that securities are correlated (rather than independent) via a common factor, and the factor risk is at least as important for fundamentals as it is for the noise in signals. To get some intuition for why Samuelson's Dictum applies under these conditions, suppose for example that the noise in the signals about the securities is close to independent, i.e., $\sigma^2_{F_\varepsilon}$ is close to zero, while $\sigma^2_{F_v}$ is relatively large. Under this condition, an informed investor can learn a lot about the true value of the factor portfolio (since the noise in the signals are close to independent across assets), but trading on this information is very risky because the factor portfolio is exposed to the common risk. Since inefficiency is the ratio of what can be learned from the signal (a lot) to what is incorporated in the price (not so much), the factor portfolio is relatively inefficient.[17] On the other hand, an informed investor can eliminate a lot of risk by holding a long-short portfolio and, to rule out aggressive trades on such portfolios, these "arbitrage portfolios" are relatively efficiently priced.

Part (c) of the proposition states that, under certain conditions, the conclusion opposite to Samuelson's Dictum obtains. This can happen, for example, if, given a fixed value of $\sigma^2_{F_v}$,

[17] We note that the measure of inefficiency is related to the Sharpe ratios of investors trading such portfolios.

the variance $\sigma^2_{F_\varepsilon}$ of the common component in the signal noise is sufficiently large. In this case, the correlated signals convey little information about the factor portfolio — indeed, in the limit as $\sigma^2_{F_\varepsilon}$ becomes infinite the inefficiency is zero: no information conveyed by the signals or prices. On the other hand, a portfolio with no factor exposure is predicted with finite noise (the most informative signal for such a portfolio has zero loading on the common signal noise), and only some of the information is impounded in the price; the inefficiency is strictly positive. In this case, learning about the factor portfolio is difficult because all signals contain correlated noise, but trading on the factor portfolio is relatively safe (because the common component in fundamental risks is comparatively small). Therefore, informed investors will make the factor portfolio relatively efficient in the sense that much of what can be learned about the factor portfolio is incorporated into the price.

Finally, part (d) of the proposition shows that the above three cases exhaust all possible scenarios, under Assumption 1. In other words, Samuelson's notion of macro vs. micro efficiency is a good one in the sense that the most and least efficient portfolios are always the factor portfolio (macro) and the arbitrage portfolios (micro), never anything in between.

3.1 Macro vs. Micro Efficiency with Many Assets

We next consider the efficiency in a market with a large number of assets. We are interested in what happens to macro and micro efficiency as the number of assets, n, grows large.

The simplest way to consider a growing number of assets is to adopt the factor structure in Assumption 1 and let $\beta_i = 1$ for any asset i.[18] In this case, all assets are symmetric and it is clear what happens to fundamentals v and noise ε as the number of assets increases — namely, there simply are more of the same type of securities. To keep the economy finite, we need to make the total supply of shares constant, which implies that the number of shares outstanding for each firm falls as $1/n$. Specifically, we let the expected supply of each security be $\bar{q}_i = \hat{q}/n$ and the supply uncertainties be $\sigma_{F_q} = \hat{\sigma}_{F_q}/n$ and $\sigma_{w_q} = \hat{\sigma}_{w_q}/n$, where $\hat{q}$, $\hat{\sigma}_{F_q}$, and $\hat{\sigma}_{w_q}$ are the corresponding (scalar) values in the economy with a single asset.

[18] A simple generalization can be made along the lines of the multi-factor extension below.

With this simple model of symmetric assets, the next proposition shows that Samuelson's Dictum always holds when n is large — that is, without relying on Assumption $1'$ (as in Proposition 3) that the factor structure in fundamentals is "strong."

In fact, the next proposition contains a stronger result: As the number of assets grows, the factor portfolio becomes the only inefficient portfolio. To appreciate this result, we note that, as we showed above — see equation (18) — the overall inefficiency η can be seen as the sum of the inefficiencies of n uncorrelated portfolios. More specifically, the overall inefficiency equals the inefficiency of the factor portfolio η^β plus $n-1$ other portfolio inefficiencies, and we show that the factor inefficiency dominates to a surprising extent:

Proposition 4 *In the symmetric single-factor model described above with $\sigma_{F_q} > 0$, the most inefficient portfolio is the factor portfolio $\beta = (1, 1, ..., 1)$ when the number of assets, n, is large enough. Further, the fraction of market inefficiency coming from the common factor approaches 100% as $n \to \infty$, that is, $\eta^\beta/\eta \to 1$.*

The fact that micro portfolios, which load exclusively on idiosyncratic shocks, are asymptotically efficient is not hard to intuit. The supply noise in such a portfolio is purely idiosyncratic, and therefore its variance goes to zero as n increases without bound. Consequently the price signal is asymptotically as informative as the fundamental signal.

The reason why the combined inefficiency of all micro portfolios tends to zero is more involved. It is a combination of the facts that (i) the variance of the (unit-norm) micro portfolios decreases as $1/n^2$ with n; (ii) the gain in the precision of the price signal, and therefore the decrease in inefficiency, from increasing n is approximately linear in this variance; and (iii) there are $n-1$ independent micro portfolios. When restricted to micro portfolios, therefore, an informed agent's utility gain over that of an uninformed one declines as $1/n$, and thus goes to zero.

On the other hand, as long as the factors are not trivial, the factor portfolio offers the informed agent a utility gain that is bounded away from zero: the supply noise in the factor does not disappear as n goes to infinity.

Multifactor-Model: The APT of Efficiency. We can generalize this result to a multi-factor model. As we shall see, in a multi-factor economy, all systematic factors have non-zero inefficiencies even with a large number of assets, but all idiosyncratic inefficiencies are eliminated, a result that "echoes" the Arbitrage Pricing Theory (APT) of Ross (1976).

To construct a model with k factors, we can reinterpret the factors F from Assumption 1 to be k-dimensional (column) vectors and, correspondingly, β is a $n \times k$ matrix of factor loadings (i.e., each column provides the loadings of that factor across assets). Further, for each vector F, we assume that the k factors are iid. Lastly, to keep the model "stationary," we assume that the eigenvalues of $\beta^\top \beta/n$ (a $k \times k$ matrix) converge to some strictly positive limits $\lambda_1, ..., \lambda_k$. Of course, we let supply decrease with n in the same way as above.

For example, we could have a two-factor model, $\beta = (\beta_1, \beta_2)$, where factor 1 is the market factor (as before), $\beta_1 = (1, 1, ..., 1)^\top$, and factor 2 is value-versus-growth, $\beta_2 = (\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, ..., \frac{(-1)^{n-1}}{2})^\top$. The specification of factor 1 means that all assets are part of the market. The alternating signs in factor 2 means that every other asset is a value stock, and the remaining ones are growth stocks. Further, the lower coefficients in factor 2 means that this factor is a weaker driver of returns. In this case, the eigenvalues of $\beta^\top \beta/n$ are 1 and $\frac{1}{4}$ for all n even.

Analogously to the APT, which describes expected returns in the presence of a factor structure, our next result characterizes asset inefficiencies. Recall that equation (1) defines the overall inefficiency η^β of the collection of factor portfolios β.

Proposition 5 *As the number of assets grows, $n \to \infty$, in the multi-factor model described above with $\sigma_{F_q} > 0$ it holds that*

[APT of Returns] *market risk premia are determined by factor loadings in the limit, that is, $E(v_i - p_i) = \sum_{j=1,...,k} \beta_{ij} \lambda_j$, where $\lambda_j \in \mathbb{R}$ is the risk premium of factor j. Consequently, a portfolio with zero loadings on all the factors has zero expected excess return.*

[APT of Efficiency] *the fraction of the market inefficiency coming from the systematic factors approaches 100%, that is, $\eta^\beta/\eta \to 1$. Consequently, a portfolio with zero loadings on all the factors has zero inefficiency.*

The standard "APT of returns" says that risk premia must be driven by systematic factors. The economics behind the APT is that, if certain assets delivered abnormal returns relative to their factor loadings, then investors could earn a return with a risk that can be diversified away — and such near-arbitrage profits are ruled out in equilibrium. Likewise, the "APT of efficiency" says that idiosyncratic inefficiencies are arbitraged away since idiosyncratic risk can be diversified away, thus leaving just the inefficiencies associated with the systematic factors when the number of securities is large. Figure 4 in the calibration section (Section 6) shows in a numerical example how the different types of inefficiency evolve as the number of securities grows large.

Proposition 5 means that, with many assets, there are two ways to try to make money on a large scale. First, one can "buy factors," that is, buy a portfolio of securities with positive factor loadings to profit from factor risk premia (the first part of the proposition). Second, one can try to exploit the inefficiency of these factors, which is sometimes called "factor timing" (the second part of the proposition). Factor timing means varying the exposure to each factor based on information on its expected return. For example, one can try to time the overall market to exploit inefficient bubbles and crashes. In contrast to buying and timing factors, some may attempt to earn near-arbitrage profits based on idiosyncratic inefficiencies, but our model predicts that such opportunities are infrequent enough that only a limited number of investors can exploit them. These predictions appear consistent with empirical evidence.[19]

Hence, with many assets, the most interesting trading opportunities occur in systematic factors, which may help explain why many investors increasingly focus their trading on exchange traded funds (ETFs), "smart beta products," futures, and other forms of factor-based investing. For example, BlackRock estimates that "the factor industry is $1.9 trillion

[19]Concerning the first part, see Roll and Ross (1980) for an early test of the APT of returns and Kelly et al. (2018) for recent evidence of factors as return drivers. Regarding the second part, tests of predictability of the market (i.e., market timing) have played a central role in the debate about market efficiency — see the literature following Campbell and Shiller (1988). For evidence of timing of other factors, see Asness et al. (2000), Greenwood and Hanson (2012), and Gupta and Kelly (2018). Regarding near-arbitrage profits, systematic evidence is rare, but it is telling that such a successful manager as Medallion Fund of Renaissance Technologies, which has reportedly consistently delivered some of the highest returns, chooses to limit its scale to the point of having no outside investors.

in AUM ... We project it will grow at a similar organic rate over the next five years, reaching $3.4 trillion by 2022."[20]

As another way to state the APT of efficiency via investors' utility, the proposition shows that informed investors can achieve almost the same utility gain by trading only factor portfolios as when they trade each individual security (when the number of securities is large). Conversely, the extra utility gain from trading a "micro" portfolio with zero factor loading is close to zero. In fact, the proposition implies an even stronger result: the total inefficiency $\sum_{j=k+1}^{n} \eta^j$ coming from the $n-k$ micro portfolios defined in (18) converges to zero as n grows (i.e., even though the number of summands increases too).[21]

Our results complement those of Glasserman and Mamaysky (2018), who also provide conditions for higher macro inefficiency. In their setting, the definition of macro vs. micro information is given exogenously, but they endogenize information choices. In contrast, we derive endogenously the meaning of macro efficiency (by looking for the most inefficient portfolio), show that Samuelson's Dictum arrives naturally as the number of securities increases, and show the potential importance of systematic factors generally, not just the overall market portfolio.

4 Falling Costs of Active and Passive Investing

Interestingly, Samuelson (1998) also conjectured how micro efficiency — but not necessarily macro efficiency — has improved over the years:

> *"The pre-1800 pattern of commercial panics had to be a case of NON MACRO-EFFICIENCY of markets. We've come a long way, baby, in two hundred years toward micro efficiency of markets: Black-Scholes option pricing, indexing of portfolio diversification, and so forth. But there is no persuasive evidence, either*

[20] "Factor Investing: 2018 Landscape," BlackRock Report, 2018.

[21] We note that total factor inefficiency equals the total inefficiency of the first k eigenvector portfolios, $\eta^\beta = \sum_{j=1}^{k} \eta^j$ defined in (18). Further, if the factor portfolios given by the columns of β are orthogonal, then these portfolios are in fact the same as the eigenvector portfolios, $\eta^{\beta_j} = \eta^j$. In this case, we can write the APT result as $\sum_{j=1}^{k} \eta^{\beta_j}/\eta \to 1$.

from economic history or avant garde theorizing, that MACRO MARKET INEFFICIENCY is trending toward extinction: The future can well witness the oldest business cycle mechanism, the South Sea Bubble, and that kind of thing. We have no theory of the putative duration of a bubble. It can always go as long again as it has already gone. You cannot make money on correcting macro inefficiencies in the price level of the stock market." [emphasis as in original]

One of the ways in which markets may have improved over time is that information costs may have come down, so it is interesting to consider whether lower information costs has the implications conjectured by Samuelson's. In particular, when Samuelson's Dictum holds for the *levels* of inefficiency (Assumption 1′ as shown in Proposition 3(b)), we can look at the relative *changes* in macro inefficiency, η^β, vs. micro inefficiency, $\eta^\perp$. Recall η^β is the inefficiency of the market portfolio, and we define micro inefficiency by $\eta^\perp := \eta^\zeta$ for any market neutral (or micro) portfolio: $\zeta^\top \beta = 0$. Interestingly, Samuelson's examples of micro efficiency appear close to our definition since option arbitrage and index arbitrage are long-short portfolios that eliminate factor risk.

Proposition 6 (Information cost and the evolution of macro vs. micro efficiency) *When the cost of information k decreases, overall asset price inefficiency η decreases and, under Assumption 1′, the macro inefficiency (η^β) decreases by more than the micro inefficiency ($\eta^\perp$) as long as γ/I is sufficiently small. Further, the number of self-directed investors remains unchanged, the numbers of informed investors I and of informed active managers M increase, the number of active investors S_a may either increase or decrease, the active management fee f_a decreases, and the passive fee f_p is unchanged.*

With the improvement in information technology, the cost of information may have decreased over time. If so, Proposition 6 shows that overall market inefficiency should have improved as a result, consistent with Samuelson's conjecture. However, Proposition 6 predicts that macro inefficiency has dropped by *more* than micro inefficiency, counter to Samuelson's conjecture. The intuition behind our result is that both macro and micro inefficiency decrease toward zero, and therefore the higher of these, macro inefficiency, must decrease by

more to reach zero. (We can only speculate regarding whether Samuelson would have considered our model "persuasive evidence" of lower macro inefficiencies based on "avant-garde theorizing.") Nevertheless, even if macro inefficiencies have decreased the most, the reamin the largest source of inefficiency, so "the oldest business cycle mechanism" may still be at play.

Another important real-world trend is that the cost of passive investing has come down over time due to low-cost index funds and exchange traded funds (ETFs). Interestingly, the cost of passive investing varies significantly across countries, giving rise to a number of cross-sectional tests, as we discuss below. First, however, we consider the model's implications for how the cost of passive investing affects security markets and the market for active asset management.

Proposition 7 (Cost of passive investing) *As the cost of passive investing $k_p = f_p$ decreases, the largest equilibrium changes as follows. The overall asset price inefficiency η increases and, under Assumption 1', the macro inefficiency (η^β) increases by more than the micro inefficiency ($\eta^\perp$) as long as γ/I is sufficiently small. Further, the number of passive investors S_p increases, the numbers investors searching for active managers S_a, self-directed investors, informed investors I, and informed active managers M decrease, the active management fee f_a may decrease or increase, and active fees in excess of passive fees $f_a - f_p$ increase.*

As seen in the proposition, we would expect that lower costs of passive investing due to index funds and ETFs should drive down the relative attractiveness of active investing and therefore reduce the amount of active investing, rendering the asset market less efficient. This effect can be visualized via Figure 2. Indeed, a reduction in the cost of passive investing implies a rise in the relative cost of active investing, corresponding to an upward shift in the dashed curve in Figure 2. In seen in the figure, such an upward shift leads to higher market inefficiency and fewer informed investors. As evidence of these predictions, Cremers et al. (2016) finds that the performance of active managers "is positively related to the market share of explicitly indexed funds [...] and negatively related to the average cost of explicit

indexing." This is consistent with Proposition 7 since higher market inefficiency naturally corresponds to better performance by active managers.

The proposition also makes predictions for fees, which we can compare with the empirical evidence. Cremers et al. (2016) find that a decline in the average fees of "indexed funds of 50 basis points ... is associated with 16 basis point lower fees charged by active funds. Overall, the results suggest that investors pay a higher price for active funds in markets in which explicitly indexed products exert less competitive pressure." In other words, active fees tend to decrease when passive fees decrease, but they move less than one-for-one so that the fee difference between active and passive fact increases when passive fees decline. Proposition 7 predicts exactly such an increase in the active-minus-passive fee difference. Our model is also consistent with a reduction in the total active fee, although this need not happen in our model. To understand this feature, note that there are two effects: First, a lower cost of passive investing directly lowers the cost of active through competitive effects as seen in equation (8). Second, having fewer active investors leads to a higher market inefficiency (η), which increases the value of active management, and hence the fee. This second effect mitigates the reduction in the active fee (and can in some cases even reverse it).

In summary, Propositions 6–9 show how changes in technology may have led to the observed "institutionalization" of the market, with a range of knock-on effects for security-market efficiency and the asset-management industry. Section 6 considers some quantitative implications.

5 Efficiency and Entropy

In this section, we seek to shed further light on the properties of market efficiency. We show (as already discussed in Section 1.3) that market efficiency is linked to the (private) economic value of information. Hence, it is natural to further explore the connection between market efficiency and information-theoretic value of information. Indeed, the idea that the economic and information-theoretic values of information are linked goes back at least to Marschak (1959), and the following proposition further establishes a link to the degree of

market inefficiency.[22,23]

Proposition 8 (Efficiency, Entropy, and the Value of Information) *In equilibrium, the following are equal:*

(a) the overall market inefficiency, η;

(b) the utility difference between informed and uninformed investors, $(u_i - u_u)\gamma$;

(c) the difference in entropy, $\text{entropy}(v|p) - \text{entropy}(v|s)$*;*

(d) the expected Kullback-Leibler divergence, KL, of the distribution of v conditional on p from that conditional on s, $E(KL)$.

The informativeness can be measured using entropy and, as is known from information theory, the entropy of a multivariate normal is a half times the log-determinant of the variance-covariance matrix (plus a constant). Therefore, the market efficiency is the difference in entropy of the distributions of fundamental values given prices, respectively given private signals. Another measure of the distance between two probability distributions is the Kullback-Leibler divergence. While the Kullback-Leibler divergence is random (since it depends on the conditioning variables p and s), the proposition establishes that the *expected* Kullback-Leibler divergence also equals the overall market inefficiency. Hence, this result establishes a new potential way to measure market efficiency and the economic value of information, namely using entropy-based methods also applied in other sciences.

[22]See also Admati and Pfleiderer (1987), who links the value of information to a ratio of determinants that coincides with our definition of inefficiency, and Cabrales et al. (2013) for a recent contribution on entropy as the economic value of information and for further references.

[23]We note that, while we state the result for the full set of assets, and associated overall market inefficiency, the natural restriction to an arbitrary set of portfolios holds.

6 Calibration and Numerical Example

By combining the equations for the active and passive fees, we see that the overall market inefficiency η can be expressed as

$$\eta = 2\gamma(f_a - f_p) + \gamma(k_p - k_a) = 2\gamma^R({f_a}^{\%} - {f_p}^{\%}) + \gamma^R({k_p}^{\%} - {k_a}^{\%}), \tag{19}$$

where $\gamma^R := \gamma W$ is the relative risk aversion, ${f_a}^{\%} := f_a/W$ and ${f_p}^{\%} := f_p/W$ are the active and passive fees as a percentage of invested wealth, and ${k_a}^{\%} := k_a/W$ and ${k_p}^{\%} := k_p/W$ are the marginal costs of active and passive management per dollar (rather than per investor).

As a simple calibration, we can set the relative risk aversion to be $\gamma^R = 3$, consider a realistic fee difference of ${f_a}^{\%} - {f_p}^{\%} = 1\%$, and assume similar marginal costs of active and passive management, ${k_p}^{\%} = {k_a}^{\%}$ (recall that active managers must additionally pay the information cost k), yielding an overall market inefficiency of $\eta = 6\%$.

This calibration derives the endogenous inefficiency based on the observed level of fees, which is itself an endogenous variable. This is similar to standard applications of the CAPM, where the expected return is derived based on the observed beta, where the beta is itself endogenous. Just as the CAPM provides useful insights on expected returns, our calculation provides an interesting implications on the magnitude of market efficiency based on the level of fees that investors are willing to pay. We can, of course, also start by choosing values for the exogenous variables and compute the equilibrium. The approximate outcomes $\eta \approx 6\%$ and ${f_a}^{\%} - {f_p}^{\%} \approx 1\%$ are actually the equilibrium outcome of the reasonable parameters used in our numerical example.[24]

[24] There are $\bar{S} = 10^8$ optimizing investors and $N = 5\times10^7$ noise allocators, each with a relative risk aversion of $\gamma^R = 3$ and wealth of $W = \$150,000$. There are $\bar{M} = 4,000$ asset managers. The fundamentals have a 2-factor structure (as in Section 3.1) where the first factor is the equal-weighted portfolio $\beta_1 = (1, 1, ..., 1)^\top$ and the second long-short factor is $\beta_2 = (b, -b, b, ..., -b)^\top$, where $b = .61$ is chosen to match the finding of Roll (1988) that "the mean R^2s were, respectively, .179 for the CAPM and .244 for the APT." The factor volatility is $0.18W(\bar{S} + N)$ and the idiosyncratic volatility is $\sigma_{w_v} = 0.4W(\bar{S} + N)$. The noise has the same factor structure, with twice the volatility of factors and idiosyncratic shocks. The number of securities is $n = 1000$, with aggregate supply $\hat{q} = 1$ and supply uncertainty $\hat{\sigma}_{F_q} = \hat{\sigma}_{w_q} = 0.14$. The marginal cost of asset management is ${k_p}^{\%} = {k_a}^{\%} = 0.10\%$ and the cost of information is $k = 3e7$. The search cost is $c = 0.6(\frac{S_a}{M})^{0.8}$. The cost of self-directed investment is zero for 20% of investors and uniformly distributed on $[0\,, 2\%]$ for the rest. Appendix A contains further information on the parametric model.

We can also try to reconcile the finding of Cremers et al. (2016) that a decline in the average fees of *"indexed funds of 50 basis points ... is associated with 16 basis point lower fees charged by active funds."* In other words, suppose that the cost ${k_p}^{\%}$ drops by 0.50%, leading to a drop in the passive fee ${f_p}^{\%}$ by the same amount and a drop in the active fee ${f_a}^{\%}$ of 0.16%. Then based on (19), we predict that overall market inefficiency increases by

$$\begin{aligned}\Delta\eta &= 2\gamma(\Delta f_a - \Delta f_p) + \gamma(\Delta k_p - \Delta k_a) \\ &= 6 \times (-0.16\% - (-0.50\%)) + 3 \times (-0.50\% - 0) = 0.54\%.\end{aligned} \tag{20}$$

In the model, the equilibrium fee and inefficiency are naturally determined jointly. Figure 3A shows how these key variables change when the cost of passive investing ${f_p}^{\%}$ varies. A change in the cost of passive investing actually does not change the active management fee in the numerical example as seen in the figure. To understand why, note first that a reduction in passive fees puts competitive pressure on active managers to lower their fees (the first term in Eqn.(8)). At the same time, however, the market becomes more inefficient since some investors move to passive management, which increases the value of active investing (the second term in Eqn.(8)). These two forces exactly offset under the specified search function, leaving the active management fee unchanged (but this not a general property of the model). As seen in the figure, if ${k_p}^{\%}$ drops by 0.50% (e.g., change of k_p from 0.60% to 0.10%) then inefficiency increases by 1.5% in the numerical example.

We can also consider how the 6% overall inefficiency is distributed between macro and micro inefficiencies as seen in Figure 4. As seen in the figure, when the number of assets is large, most inefficiency arises from macro sources. In fact, at the right end of the figure with 1000 assets, we see that 81% of the overall inefficiency is due to the inefficiency of the expected market portfolio $\beta_1 = (1, 1, ..., 1)^\top$, 18% is due to the other systematic factor, namely the relative-value portfolio $\beta_2 = (.61, -.61, ..., -.61)^\top$, and the remaining 1% is due to all the 998 micro portfolios. The low degree of inefficiency stemming from the micro portfolios may seem shocking, but it arises from investors' ability to diversify such risk when there are as many as 1000 securities in our example. In other words, micro inefficiencies

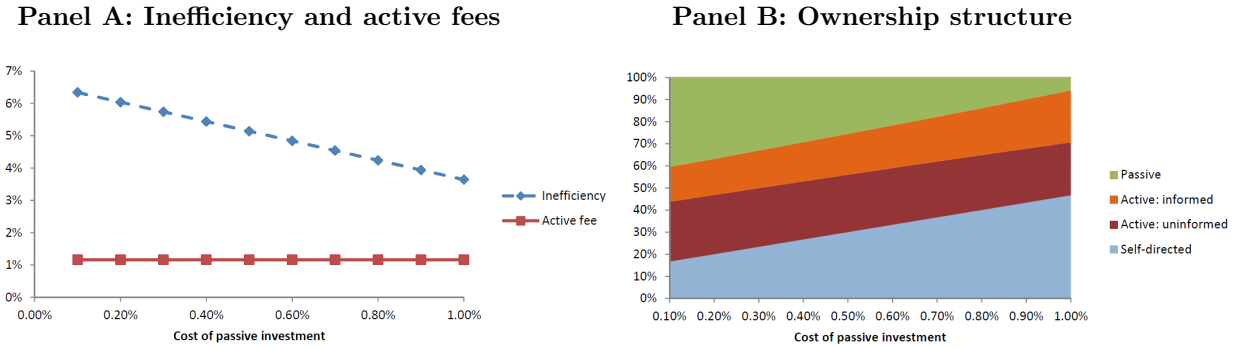

Figure 3: **Numerical Example.** This figure shows properties of the model implied by different values of the percentage cost of passive investment, $f_p^{\%}$ (listed on the x-axis). Panel A shows the inefficiency η and the active fee $f_a^{\%}$. Panel B shows shows the fraction of ownership that is active management (informed I and uninformed $N\frac{\bar{M}-M}{\bar{M}}$), passive management ($S_p$), and self-directed.

are diminished when informed investing virtually eliminates near-arbitrage opportunities, at least in the model. Hence, most of the inefficiency arises from the non-diversifiable risk due to the two factors. Most of the inefficiency is in the expected market portfolio, but also a non-trivial part is in the second factor, which is a long-short portfolio such as the high-minus-low (HML) value factor or the small-minus-big (SMB) size factor used in much of empirical finance (see Fama and French (1993)). Hence, non-trivial mispricing can exist when many inefficient trades are correlated, consistent with the empirical evidence that most return drivers indeed are based on factor structures of such correlated trades (see Kelly et al. 2018 and references therein). In contrast, truly idiosyncratic mispricing should be minimal according to the model — quantitative predictions that may or not stand the test of data, especially before transaction costs.[25]

Figure 3B shows how the ownership structure depends on the passive fees in the numerical example. We see that lower passive costs imply more passive asset management, less active investment, especially informed active investment, and less self-directed investment.

These findings can be viewed as the model-based counterpart to the recent trends in real-

[25]Gupta and Kelly (2018) find that many factors can be timed, not just the market, which likely poses a challenge to the estimate that as much as 81% of the inefficiency stems from the market portfolio. However, their setting includes many more factors than the study of Roll (1988) used for our choice of parameters (see Footnote 24) so a real test of the model should use parameters consistent with the test portfolios.

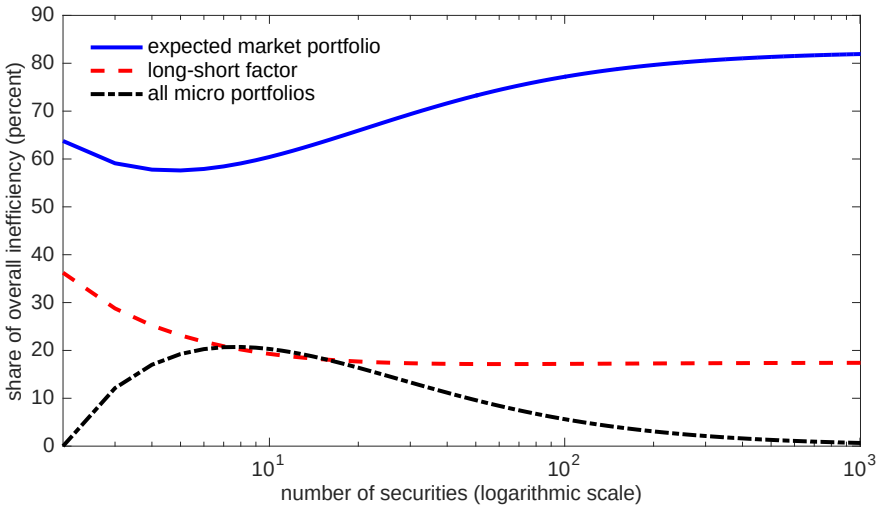

Figure 4: **Decomposing Overall Inefficiency.** This figure shows the share of the overall market inefficiency arising from the expected market portfolio (solid line), long-short portfolios such as the Fama and French (1993) factors called high-minus-low (HML) and small-minus-big (SMB) used in much of empirical finance (dashed line), and the sum of all micro portfolios. With many assets, the overall inefficiency is mostly due to the former two kinds of inefficiency, both macro in nature, consistent with Samuelson's Dictum.

world markets seen in Figure 1. Over the past two decades (moving left-to-right in Figure 1), we have seen an increase in passive management and a decrease in active management and self-directed investment, consistent with the model if we assume that the cost of passive investment have declined (moving right-to-left in Figure 3B).

7 Conclusion and Testable Implications

We model how investors choose between active and passive management, how active and passive managers choose their portfolios, and how security prices are set. We provide a theoretical foundation for Samuelson's Dictum by showing that macro inefficiency is greater than micro inefficiencies under realistic conditions. We calibrate the central economic magnitudes, thus providing a potential explanation for the recent trends in asset management and financial markets.

Our model provides new testable implications to be explored in future empirical research.

First, the model provides a clear link between active fees and market efficiency; in our calibration, overall market efficiency is six times the cost of active management.

Second, the model shows how to decompose this overall inefficiency into the inefficiency of the market portfolio, the inefficiency of other factor portfolios, and that of truly idiosyncratic micro bets. In particular, the fraction of variance explained by each of these types of returns should be linked to the Sharpe ratios that can achieved by trading them.

Third, the model makes predictions on the impact of the ongoing, widespread reductions in the cost of passive management on capital markets and the industrial organization of the asset-management industry. In particular, we show that falling fees of passive investing will increase market inefficiency, lower active fees by less than the passive fees, lower the fraction of active investors, lower the number of active managers, and increase the fraction of uninformed active managers (i.e., closet indexers).

Fourth, the model has implications for the optimal informed and uninformed trading strategies. Since passive funds and indexes have incentives to mimic the optimal uninformed strategy, these results can be seen as predictions for which types of indices should emerge as the most successful, and likewise for the active strategies.

Fifth, we show how market inefficiency is linked to information entropy. Hence, it might be possible to adapt entropy methods used in other sciences to estimate market efficiency and further test the model.

References

Admati, A. R. (1985). A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica*, 629–657.

Admati, A. R. and P. Pfleiderer (1987). Viable allocations of information in financial markets. *Journal of Economic Theory 43*(1), 76–115.

Admati, A. R. and P. Pfleiderer (1990). Direct and indirect sale of information. *Econometrica 58*, 901–928.

Asness, C., A. Frazzini, and L. H. Pedersen (2018). Quality minus junk. *Review of Accounting Studies, forthcoming*.

Asness, C., T. Moskowitz, and L. H. Pedersen (2013). Value and momentum everywhere. *The Journal of Finance 68*(3), 929–985.

Asness, C. S., J. A. Friedman, R. J. Krail, and J. M. Liew (2000). Style timing: Value versus growth. *Journal of Portfolio Management 26*(3), 50–60.

Berk, J. B. and J. H. V. Binsbergen (2015). Measuring skill in the mutual fund industry. *Journal of Financial Economics 118*(1), 1–20.

Berk, J. B. and R. C. Green (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy 112*(6), 1269–1295.

Breugem, M. and A. Buss (2018). Institutional investors and information acquisition: Implications for asset prices and informational efficiency. *The Review of Financial Studies 32*(6), 2260–2301.

Buffa, A. M., D. Vayanos, and P. Woolley (2014). Asset management contracts and equilibrium prices. *Boston University, working paper*.

Cabrales, A., O. Gossner, and R. Serrano (2013). Entropy and the value of information for investors. *The American Economic Review 103*(1), 360–377.

Campbell, J. Y. and R. J. Shiller (1988). The dividend-price ratio and expectations of future dividends and discount factors. *The Review of Financial Studies 1*(3), 195–228.

Cremers, M., M. A. Ferreira, P. Matos, and L. Starks (2016). Indexing and active fund management: International evidence. *Journal of Financial Economics 120*(3), 539–560.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2017). International tests of a five-factor asset pricing model. *Journal of Financial Economics 123*(3), 441–463.

Frazzini, A., D. Kabiller, and L. H. Pedersen (2018). Buffett's alpha. *Financial Analysts Journal, forthcoming*.

French, K. R. (2008). Presidential address: The cost of active investing. *The Journal of Finance 63*(4), 1537–1573.

García, D. and J. M. Vanden (2009). Information acquisition and mutual funds. *Journal of Economic Theory 144*(5), 1965–1995.

Gârleanu, N. and L. H. Pedersen (2018). Efficiently inefficient markets for assets and asset management. *The Journal of Finance 73*(4), 1663–1712.

Gennaioli, N., A. Shleifer, and R. Vishny (2015). Money doctors. *The Journal oif Finance 70*(1), 91–114.

Glasserman, P. and H. Mamaysky (2018). Investor information choice with macro and micro information. *Columbia University, working paper*.

Graham, B. and D. L. Dodd (1934). *Security Analysis*. McGraw-Hill.

Greenwood, R. and S. G. Hanson (2012). Share issuance and factor timing. *The Journal of Finance 67*(2), 761–798.

Grossman, S. J. (1995). Dynamic asset allocation and the informational efficiency of markets. *The Journal of Finance 50*(3), 773–787.

Grossman, S. J. and J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *American Economic Review 70*, 393–408.

Gruber, M. J. (1996). Another puzzle: The growth in actively managed mutual funds. *The Journal of Finance 51*(3), 783–810.

Gupta, T. and B. T. Kelly (2018). Factor momentum everywhere. *Available at SSRN 3300728*.

Jung, J. and R. J. Shiller (2005). Samuelson's dictum and the stock market. *Economic Inquiry 43*(2), 221–228.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp (2016). A rational theory of mutual funds' attention allocation. *Econometrica 84*(2), 571–626.

Kacperczyk, M. T., J. B. Nosal, and S. Sundaresan (2018). Market power and price informativeness. *Available at SSRN 3137803*.

Kelly, B., S. Pruitt, and Y. Su (2018). Characteristics are covariances: A unified model of risk and return. *National Bureau of Economic Research, working paper*.

Marschak, J. (1959). Remarks on the economics of information. *In Contributions to Scientific Research in Management, 7998. Los Angeles: University of California, Western Data Processing Center*.

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

Pastor, L., R. F. Stambaugh, and L. A. Taylor (2015). Scale and skill in active management. *Journal of Financial Economics 116*(1), 23–45.

Pedersen, L. H. (2018). Sharpening the arithmetic of active management. *Financial Analysts Journal 74*(1), 21–36.

Petajisto, A. (2009). Why do demand curves for stocks slope down? *Journal of Financial and Quantitative Analysis 44*(5), 1013–1044.

Roll, R. (1977). A critique of the asset pricing theory's tests part i: On past and potential testability of the theory. *Journal of financial economics 4*(2), 129–176.

Roll, R. (1988). R-squared. *Journal of Finance 43*(2), 541–566.

Roll, R. and S. A. Ross (1980). An empirical investigation of the arbitrage pricing theory. *The Journal of Finance 35*(5), 1073–1103.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory 13*, 341–360.

Samuelson, P. A. (1998). Summing upon business cycles: Opening address. In J. C. Fuhrer and S. Schuh (Eds.), *Beyond Shocks: What Causes Business Cycles*, pp. 33–36. Boston, MA: Federal Reserve Bank of Boston.

Shiller, R. J. (2001). *Irrational Exuberance, 2nd ed.* New York, NY: Broadway Books.

Stambaugh, R. F. (2014). Presidential address: Investment noise and trends. *Journal of Finance 69*, 1415–1453.

Stein, J. C. (2009). Presidential address: Sophisticated investors and market efficiency. *The Journal of Finance 64*(4), 1517–1548.

Van Nieuwerburgh, S. and L. Veldkamp (2010). Information acquisition and under-diversification. *The Review of Economic Studies 77*(2), 779–805.

Vayanos, D. and P. Woolley (2013). An institutional theory of momentum and reversal. *Review of Financial Studies 26*, 1087–1145.

Veldkamp, L. (2011). *Information choice in macroeconomics and finance*. Princeton University Press.

A Appendix: Further Analysis and Proofs

A.1 Calibration: parametric example with closed-form solution.

In our calibration, we consider the following specification of investors' search cost:

$$c\left(M, S_a\right) = \bar{c}\left(\frac{S_a}{M}\right)^{\alpha} \text{ for } M > 0 \quad \text{and} \quad c(M, S_a) = \infty \text{ for } M = 0, \tag{A.1}$$

where $\alpha > 0$ and $\bar{c} > 0$ are parameters. Combining this with (9)–(11) gives

$$\eta = 2\gamma\left(\frac{k_a - k_p}{2} + (\bar{c}k^{\alpha})^{\frac{1}{1+\alpha}}\right), \tag{A.2}$$

which shows how market inefficiency depends on search costs $(\bar{c}, \alpha)$, asset management costs (k_a, k_p), and information costs (k).

A.2 Notation.

In the proofs, we will use the following notation for any random variables x and y, $\Sigma_x := \text{var}(x)$ and $\Sigma_{x|y} := \text{var}(x|y)$. In addition, we define the functions

$$g^I(I, M) = c(M, I - NM/\bar{M}) - \frac{\eta(I)}{2\gamma} - \frac{k_p - k_a}{2} \tag{A.3}$$

$$g^M(I, M) = \frac{M}{I - N\frac{M}{\bar{M}}}k - \frac{\eta(I)}{2\gamma} - \frac{k_p - k_a}{2}. \tag{A.4}$$

Given (9), (11), and the definition of I, at any interior equilibrium we have $g^I(I, M) = 0$ and $g^M(I, M) = 0$.

A.3 Deriving an equilibrium.

Here we review a few of the details of the Grossman and Stiglitz (1980) logic, which determines the asset-market equilibrium. We explained the investors' choices and managers' entry decisions in the body of the paper.

An agent having conditional expectation of the final value μ and variance V optimally demands a number of shares equal to

$$x = (\gamma V)^{-1}(\mu - p). \tag{A.5}$$

To compute the relevant expectations and variance, we conjecture the form (4) for the price and introduce a slightly simpler "auxilary" price, $\hat{p} = v - \bar{v} + \varepsilon - \theta_q(q - \bar{q})$, with the same

information content as p:

$$\mathrm{E}[v|p] = \mathrm{E}[v|\hat{p}] = \bar{v} + \beta_{v,\hat{p}}\hat{p} = \bar{v} + \Sigma_v \Sigma_{\hat{p}}^{-1} \hat{p} \tag{A.6}$$

$$\mathrm{E}[v|s] = \mathrm{E}[v|v+\varepsilon] = \bar{v} + \beta_{v,s}(s - \bar{v}) = \bar{v} + \Sigma_v \Sigma_s^{-1}(s - \bar{v}) \tag{A.7}$$

$$\Sigma_{v|p} = \Sigma_{v|\hat{p}} = \left(\Sigma_v^{-1} + \Sigma_{\varepsilon+\theta_q q}^{-1}\right)^{-1} = \Sigma_v \left(\Sigma_v + \Sigma_{\varepsilon+\theta_q q}\right)^{-1} \Sigma_{\varepsilon+\theta_q q} \tag{A.8}$$

$$\Sigma_{v|s} = \left(\Sigma_v^{-1} + \Sigma_\varepsilon^{-1}\right)^{-1} = \Sigma_v \left(\Sigma_v + \Sigma_\varepsilon\right)^{-1} \Sigma_\varepsilon. \tag{A.9}$$

We can now insert these demands into the market-clearing condition (6), which is a linear equation in the random variables q and s. Given that this equation must hold for all values of q and s, the aggregate coefficients on these variables must equal zero, and similarly, the constant term must be zero. Solving these three equations leads to the coefficients in the price function (4):

$$\theta_0 = \bar{v} - \gamma \left(U \left(\Sigma_{v|p}\right)^{-1} + I \left(\Sigma_{v|s}\right)^{-1}\right)^{-1} \bar{q} \tag{A.10}$$

$$\theta_q = \frac{\gamma}{I} \Sigma_\varepsilon \tag{A.11}$$

$$\theta_s = \left(U \left(\gamma\Sigma_{v|p}\right)^{-1} + I \left(\gamma\Sigma_{v|s}\right)^{-1}\right)^{-1} \left(\theta_q^{-1} + U \left(\gamma\Sigma_{v|p}\right)^{-1} \Sigma_v \Sigma_{\hat{p}}^{-1}\right). \tag{A.12}$$

A.4 Further Comparative Statics

Section 4 considers comparative statics with respect to some key changes in the market, namely the costs of active and passive investing. Another potential change resulting from the rise in delegated management is a reduction in "noise trading." While we have emphasized that supply uncertainty arises for several rational reasons (e.g., firms issuing or repurchasing shares), it can also arise simply from investors making irrational trades. If this so-called noise trading is due primarily to individuals, then it may have gone down over time as emphasized by Stambaugh (2014). We can also consider the implications of a change in noise in the context of our model.

Proposition 9 (Change in noise) *Suppose that the variance of the supply noise is $\Sigma_q = z\bar{\Sigma}_q$, where z is a scalar. Then a lower z (i.e., a lower supply uncertainty) results in lower S_a, M, and I, higher S_p, while the overall asset price inefficiency η may either increase or decrease.*

We see that a reduction in noise trading should lead to a reduction in the number of active investors and active managers. The effect on market inefficiency is ambiguous — it depends on how many patsies (noise traders) vs. sharks (informed investors) have left the "poker table."

A.5 Proofs

A number of the results that we prove below can be derived quite quickly based on the following lemma. To accommodate all the related propositions, we allow for the generality required by Proposition 5, which we spell out as a formal assumption.

Assumption 1″ *Fundamentals have a factor structure:*

$$v = \bar{v} + \beta F_v + w_v \tag{A.13}$$
$$\varepsilon = \beta F_\varepsilon + w_\varepsilon \tag{A.14}$$
$$q = \bar{q} + \beta F_q + w_q, \tag{A.15}$$

where $\bar{v}, \bar{q} \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{n\times k}$, the common factors F_v, F_ε, and F_q are k-dimensional random iid variables with zero means and variances of each entry $\sigma^2_{F_v}$, $\sigma^2_{F_\varepsilon}$, and $\sigma^2_{F_q}$, respectively, and the idiosyncratic shocks w_v, w_ε, and w_q taking value in $\mathbb{R}^n$ are i.i.d. across assets with variances $\sigma^2_{w_v}$, $\sigma^2_{w_\varepsilon}$, respectively $\sigma^2_{w_q}$ for each asset.

We use the notation

$$G = \Sigma_{v|s}^{-\frac{1}{2}} \Sigma_{v|p} \Sigma_{v|s}^{-\frac{1}{2}} \tag{A.16}$$
$$O = \beta\beta^\top. \tag{A.17}$$

Lemma 1 *Under Assumption 1″, the matrix G has the same eigenvectors as O. Letting λ be an eigenvalue of O, the eigenvalue of G corresponding to the same eigenvector is given by*

$$g(\lambda) \equiv 1 + \frac{X(\lambda)Y(\lambda)}{1 + X(\lambda) + Y(\lambda)}, \tag{A.18}$$

with

$$X(\lambda) \equiv \frac{\sigma^2_{w_v} + \sigma^2_{F_v}\lambda}{\sigma^2_{w_\varepsilon} + \sigma^2_{F_\varepsilon}\lambda} \tag{A.19}$$
$$Y(\lambda) \equiv \gamma^2 I^{-2} \left(\sigma^2_{w_\varepsilon} + \sigma^2_{F_\varepsilon}\lambda\right)\left(\sigma^2_{w_q} + \sigma^2_{F_q}\lambda\right). \tag{A.20}$$

If Assumption 1′ holds, then g is an increasing function.

Proof of Lemma 1. We start by noting that the variance matrices of the fundamental quantities v, ε, and q have the form $aI_n + bO$ for appropriate (positive) scalars a and b. E.g.,

$$\Sigma_v = \sigma^2_{w_v} I_n + \sigma^2_{F_v} O. \tag{A.21}$$

We need to work with a slightly more general set of matrices. Specifically, with $B = \beta^\top\beta$, we consider matrices of the form $aI_n + \beta f(\beta^\top\beta)\beta^\top$, where $f : \mathbb{R}_+ \to \mathbb{R}$ is a finite-valued function.[26]

[26]For a symmetric positive definite matrix $S = UDU^{-1}$ with D diagonal, $f(S) \equiv Uf(D)U^{-1}$, with $f(D)$ a diagonal matrix of values of f. For our purposes, we can treat $f(S)$ merely as notation for $Uf(D)U^{-1}$.

It is clear that the set of matrices of the above form is closed under the arithmetic operations of addition, subtraction, and multiplication. Further, for any function f such that $f(x)x + 1 \neq 0 \ \forall x \in \mathbb{R}_+$, we have

$$\left(I_n + \beta f(\beta^\top \beta)\beta^\top\right)^{-1} = I_n + \beta \hat{f}(\beta^\top \beta)\beta^\top \tag{A.22}$$

$$= I_n - \beta(I_n + \beta^\top \beta f(\beta^\top \beta))^{-1} f(\beta^\top \beta)\beta^\top, \tag{A.23}$$

with

$$\hat{f}(x) = -\frac{f(x)}{1 + xf(x)} \tag{A.24}$$

also satisfying $\hat{f}(x)x + 1 \neq 0$.

It follows that all variance-covariance matrices, their inverses, as well as any other matrices describing the equilibrium, have the form $aI_n + \beta f(B)\beta^\top$. (Note that, since all matrices that have to be inverted are known to be invertible, it is never the case that $f(x)x = -1$.) It is immediately apparent that the eigenvectors of $aI_n + \beta f(B)\beta^\top$ are the k eigenvectors of $\beta\beta^\top$ that are not associated with zero eigenvalues (equivalently, they equal βy with y eigenvector of B), and $n - k$ linearly independent vectors orthogonal to the columns of β. For the first type of eigenvector, given $\beta\beta^\top y = \lambda y$, the associated eigenvalue is $1 + \lambda f(\lambda)$; for the second it is 1.

Moreover, given two such matrices $M_1 = a_1 I_n + \beta f_1(B)\beta^\top$ and $M_2 = a_2 I_n + \beta f_2(B)\beta^\top$ and the bivariate function $F(x, y)$ being either addition, subtraction, multiplication, or division, the eigenvalue of $F(M_1, M_2)$ corresponding to eigenvalue λ of O equals $F(a_1 + \lambda f_1(\lambda), a_2 + \lambda f_2(\lambda))$.

Consider now the matrix

$$G = \Sigma_{v|s}^{-\frac{1}{2}} \Sigma_{v|p} \Sigma_{v|s}^{-\frac{1}{2}} = \Sigma_{v|s}^{-1} \Sigma_{v|p} = \left(\Sigma_v^{-1} + \Sigma_\varepsilon^{-1}\right)\left(\Sigma_v^{-1} + (\Sigma_\varepsilon + \theta_q \Sigma_q \theta_q)^{-1}\right)^{-1}. \tag{A.25}$$

All its eigenvectors are described above. Its eigenvalue associated with any O eigenvector is a function of the corresponding eigenvalue of O given by (A.25). Specifically, we have

$$g(\lambda) = \frac{\left(\sigma_{w_v}^2 + \sigma_{F_v}^2 \lambda\right)^{-1} + \left(\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2 \lambda\right)^{-1}}{\left(\sigma_{w_v}^2 + \sigma_{F_v}^2 \lambda\right)^{-1} + \left(\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2 \lambda + \gamma^2 I^{-2}\left(\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2 \lambda\right)^2 \left(\sigma_{w_q}^2 + \sigma_{F_q}^2 \lambda\right)\right)^{-1}}. \tag{A.26}$$

It is a simple manipulation to check that equation (A.18) holds.

Further, it is also immediate that the right-hand side of that equation increases in λ as long as the positive quantities X and Y do. It is clear that Y is increasing, while X increases if and only if Assumption 1′ holds. ■

Proof of Proposition 1. Part 1. (a) Start with the market clearing condition

$$\begin{aligned} q &= Ux_u + Ix_i \\ &= U\left(\gamma\Sigma_{v|p}\right)^{-1}\left(\mathrm{E}[v|p] - p\right) + I\left(\gamma\Sigma_{v|s}\right)^{-1}\left(\mathrm{E}[v|s,p] - p\right). \end{aligned} \tag{A.27}$$

Take expectations conditional on p and rewrite to get

$$\mathrm{E}[q|p] = \left(U + I\Sigma_{v|s}^{-1}\Sigma_{v|p}\right)x_u. \tag{A.28}$$

Solving for x_u yields equation (15).

(b) Let's use the notation $A \sim B$ for two matrices that are scalar multiplies of each other. To see the implications of the sufficient condition $\Sigma_v \sim \Sigma_\varepsilon \sim \Sigma_q^{-1}$, we note the following.

$$\theta_q \sim \Sigma_\varepsilon \tag{A.29}$$

$$\theta_s \sim I_n \tag{A.30}$$

$$\Sigma_{v|s} = \Sigma_v\left(\Sigma_v + \Sigma_\varepsilon\right)^{-1}\Sigma_\varepsilon \sim \Sigma_\varepsilon \tag{A.31}$$

$$\Sigma_{v|p} = \Sigma_v\left(\Sigma_v + \Sigma_\varepsilon + \theta_q\Sigma_q\theta_q\right)^{-1}\left(\Sigma_\varepsilon + \theta_q\Sigma_q\theta_q\right) \sim \Sigma_\varepsilon \tag{A.32}$$

Consequently, $\Sigma_{v|s}^{-1}\Sigma_{v|p}$ is a scalar and x_u is proportional to $\mathrm{E}[q|p]$.

(c) Under Assumption 1, $\Sigma_{v|s}$ and $\Sigma_{v|p}$ commute, which implies that $\Sigma_{v|s}^{-1}\Sigma_{v|p}$ is positive definite, and therefore $(U + I\Sigma_{v|s}^{-1}\Sigma_{v|p})^{-1}$ is positive definite.

Further, the assumptions of Lemma 1 are satisfied. It follows that $H \equiv (U + I\Sigma_{v|s}^{-1}\Sigma_{v|p})^{-1}$ takes the form $a_0 - a_1 O$. Equivalently, that the function h giving the eigenvalues $h(\lambda)$ of H be decreasing, which is itself equivalent with the function g giving the eigenvalues of G being increasing. Assumption $1'$ is sufficient for this conclusion.

(b) This result follows from equation (16), taking into account the sign of A_1 and the normalization of β.

(c) Under the assumptions stated, the investments in each asset are as in a single-asset Grossman-Stiglitz world. Since $\Sigma_{v|p}$ increases with Σ_q, the coefficient A — which is trivially a scalar in a one-asset world — decreases with the variance of q_i. ∎

Proof of Proposition 2. (a) Consider demands conditional on realized supply:

$$\begin{aligned} \mathrm{E}[x_i|q] &= \left(\gamma\Sigma_{v|s}\right)^{-1}\left(\bar{v} - \mathrm{E}[p|q]\right) \\ &= \left(\gamma\Sigma_{v|s}\right)^{-1}\pi + \left(\gamma\Sigma_{v|s}\right)^{-1}\theta_s\theta_q(q - \bar{q}) \end{aligned} \tag{A.33}$$

$$\begin{aligned} \mathrm{E}[x_u|q] &= \left(\gamma\Sigma_{v|p}\right)^{-1}\left(\mathrm{E}[\mathrm{E}[v|p]|q] - \mathrm{E}[p|q]\right) \\ &= \left(\gamma\Sigma_{v|p}\right)^{-1}\pi + \left(\gamma\Sigma_{v|p}\right)^{-1}\left(\theta_s - \Sigma_v\Sigma_{\hat{p}}^{-1}\right)\theta_q(q - \bar{q}), \end{aligned} \tag{A.34}$$

where π is the risk premium

$$\pi = \left(U \left(\gamma \Sigma_{v|p} \right)^{-1} + I \left(\gamma \Sigma_{v|s} \right)^{-1} \right)^{-1} \bar{q}.$$

Each of the following inequalities holds as long as all matrices involved commute with each other, which is the case both under Assumption 1 and under Assumption 2.

$$\theta_s > \theta_s - \Sigma_v \Sigma_{\hat{p}}^{-1} \tag{A.35}$$

$$\theta_s \theta_q > \left(\theta_s - \Sigma_v \Sigma_{\hat{p}}^{-1} \right) \theta_q \tag{A.36}$$

$$\left(\gamma \Sigma_{v|s} \right)^{-1} \theta_s \theta_q > \left(\gamma \Sigma_{v|p} \right)^{-1} \left(\theta_s - \Sigma_v \Sigma_{\hat{p}}^{-1} \right) \theta_q. \tag{A.37}$$

The desired conclusion follows, both because uninformed investors update their estimate of the value, which mitigates the direct impact of the price on their demand, and because they face more risk.

(b) We have

$$\begin{aligned} \mathrm{E}[x_i|s] &= \left(\gamma \Sigma_{v|s} \right)^{-1} \left(\mathrm{E}[v|s] - \mathrm{E}[p|s] \right) \\ &= \left(\gamma \Sigma_{v|s} \right)^{-1} \pi + \left(\gamma \Sigma_{v|s} \right)^{-1} \left(\Sigma_v \Sigma_s^{-1} - \theta_s \right) (s - \bar{v}) \end{aligned} \tag{A.38}$$

$$\begin{aligned} \mathrm{E}[x_u|s] &= \left(\gamma \Sigma_{v|p} \right)^{-1} \left(\mathrm{E}[\mathrm{E}[v|p]|s] - \mathrm{E}[p|s] \right) \\ &= \left(\gamma \Sigma_{v|p} \right)^{-1} \pi + \left(\gamma \Sigma_{v|p} \right)^{-1} \left(\Sigma_v \Sigma_{\hat{p}}^{-1} - \theta_s \right) (s - \bar{v}). \end{aligned} \tag{A.39}$$

Under Assumption 1 or Assumption 2, $\Sigma_v \Sigma_{\hat{p}}^{-1} < \theta_s < \Sigma_v \Sigma_s^{-1}$ (see equation (A.12)). (In general, due to the market clearing, $I\mathrm{E}[x_i|s] + U\mathrm{E}[x_u|s] = \bar{q}$.) The conclusion follows. ■

Proof of Proposition 3. (a) Under Assumption 2, G is a scalar (see proof of Proposition 1, part 1a), and thus all portfolios have the same inefficiency.

(b) We note that the market portfolio, β, is the only non-zero eigenvector of O. It is the maximum-inefficiency portfolio if and only if the associated eigenvalue of G is higher than for the other eigenvectors of O — the ones with O-eigenvalues zero. It is sufficient, then, that the function g in Lemma 1 be increasing, which obtains under Assumption 1′.

(c) This can be shown via numerical example, or by fixing all other parameters and observing that g is a decreasing function when $\sigma^2_{F_\varepsilon}$ is large enough. (It is perhaps easier to see that $(g(\lambda) - 1)^{-1}$ is increasing over a fixed domain when $\sigma^2_{F_\varepsilon}$ is large enough.) ■

Proof of Proposition 4. This proposition follows from Lemma 1 by letting n be large. One point to keep in mind is that the equilibrium mass of agents investing with informed managers, I, also depends on n. It suffices for our purposes here, however, that this quantity be bounded — above and away from zero. In fact, a stronger statement holds, under our mild regularity assumptions: I has a strictly positive limit with n. This is the mass of informed agents that arises as equilibrium when the inefficiency, as a function of I, is given

by

$$\eta_\infty(I) := \frac{1}{2}\log\left(1 + \frac{X_\infty Y_\infty}{1 + X_\infty + Y_\infty}\right), \tag{A.40}$$

where $X_\infty := \lim_{n\to\infty} X(n)$ and $Y_\infty := \lim_{n\to\infty} Y(n;n)$, with $Y(\lambda;n)$ notation for $Y(\lambda)$ defined in (A.19)–(A.20) that makes explicit the dependence of the variance $\sigma^2_{F_q}$ on n. We note that this is the efficiency obtaining in an economy with an exact factor structure, i.e., one where $\sigma^2_{F_v} = \sigma^2_{F_\varepsilon} = \sigma^2_{F_q} = 0$ — effectively, a one-asset economy where the risks are captured by what, with multiple assets, are the factors.

We show below that, in fact, the sequence η_n tends to η_∞ pointwise. Furthermore, solving for M as a function of I via (A.4) — g^M increases strictly with M — finding an equilibrium, for every n, comes down to solving one equation, namely (A.3), in one unknown, namely I. Assuming that an interior equilibrium to the "limit economy" exists, and denoting by I_∞ the associated mass of agents investing based on information, we have that, for any sufficiently small $\delta > 0$, for all n large enough the right-hand side of (A.3) takes opposite signs at $I_\infty - \delta$ and $I_\infty + \delta$, and thus has a zero in $(I_\infty - \delta, I_\infty + \delta)$.[27]

All that remains is to take a closer look at

$$\lim_{n\to\infty} \eta_n(I) = \lim_{n\to\infty}\left(\frac{1}{2}\log(g(n;n)) + \frac{1}{2}(n-1)\log(g(0;n))\right), \tag{A.41}$$

where we write $g(\lambda;n)$ for the eigenvalue $g(\lambda)$ from (A.18).

The first term clearly tends to $\eta_\infty(I)$. We note that both X_∞ and Y_∞ are positive constants (as long as $\sigma^2_{F_v} > 0$). The second term, which equals the total inefficiency due to the "micro" portfolios, is of order n^{-1}, since $Y(0;n)$ is of order n^{-2}. ∎

Proof of Proposition 5. We adapt the proof of Proposition 4 in a straightforward way. As noted there, the sum $\frac{1}{2}\sum_{i>k}\log(g_i)$ is of the order n^{-1}, while each of the other eigenvalues represents an inefficiency level that is non-zero in the limit. ∎

Proof of Proposition 6. As noted above, at any interior equilibrium, $g^I(I,M) = 0$ and $g^M(I,M) = 0$. Further, equation (A.4) defines implicitly a function $\mathcal{M}(I)$. Given the regularity assumptions on c, equation (A.3) defines a function $\mathcal{I}(M)$, and this function is strictly increasing. We complete the definition of these functions by imposing the population constraints $\mathcal{M} \le \bar{M}$ and $\mathcal{I}(M) \le S_a + S_p + N\frac{M}{\bar{M}}$, noting that $S_a + S_p$ is determined only by $\bar{S}$, k_p, and the distribution of d.

It is easy to analyze the corner equilibria, and in particular the case $M = \bar{M}$. (We note that, if I is maximal, so are S_a and M.) Suppose then that $M < \bar{M}$, meaning that $g^M(\mathcal{I}(\bar{M}), \bar{M}) > 0$. On the other hand, by definition of $\mathcal{I}$, $g^I(\mathcal{I}(\bar{M}), \bar{M}) \le 0$. Given that

[27] For our purposes, all that matters is that one such value of δ exists with $\delta < I_\infty$.

both of these functions are continuous in M, and equal zero at M, we deduce

$$\frac{d}{dM}\left(g^M(\mathcal{I}(M),M)-g^I(\mathcal{I}(M),M)\right)>0. \tag{A.42}$$

Using that $\mathcal{I}(M)$ is interior, we compute the derivative $\mathcal{I}'(M)$ and plug it in the equation above to conclude

$$g^M_M-g^M_I\frac{g^I_M}{g^I_I}>0 \tag{A.43}$$

in a neighborhood of the equilibrium value M. Noting that $g^I_I>0$, we have

$$g^I_M g^M_I < g^I_I g^M_M. \tag{A.44}$$

Consider now the effect of decreasing k. The dependence of I and M on k is given as a solution to

$$\begin{pmatrix} g^M_I & g^M_M \\ g^I_I & g^I_M \end{pmatrix}\begin{pmatrix} I_k \\ M_k \end{pmatrix} = -\frac{M}{S_a}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{A.45}$$

and therefore by

$$\begin{pmatrix} I_k \\ M_k \end{pmatrix} = \frac{1}{g^I_M g^M_I - g^I_I g^M_M}\begin{pmatrix} g^I_M \\ -g^I_I \end{pmatrix}\left(-\frac{M}{S_a}\right). \tag{A.46}$$

Given $g^I_I>0$, $g^I_M<0$, and $g^I_M g^M_I - g^I_I g^M_M<0$ from (A.44), we conclude $I_k<0$ and $M_k<0$. A decrease in k, therefore, translates into a higher I and into a higher M. It is easy to see that S_a may either increase or decrease; in particular, it may decrease if $c_M\approx 0$ and $|c_{S_a}|$ is very large around the equilibrium.

It is clear that the number of self-directed investors is unaffected, f_a decreases from equation (8) if η does, and $f_p=k_p$ is unaffected.

To see that the higher I leads to a lower inefficiency η, we note that the matrix $\Sigma_{v|p}$ decreases — in the operator sense, i.e., as a quadratic form — with I, because

$$\Sigma_{v|p}=\Sigma_v-\Sigma_v\left(\Sigma_v+\Sigma_\varepsilon+\theta_q\Sigma_q\theta_q\right)^{-1}\Sigma_v \tag{A.47}$$

and $\theta_q=\frac{\gamma}{I}\Sigma_\varepsilon$. (The one non-obvious mathematical fact that needs to be used is that, for A and B symmetric matrices, $A\geq B\geq 0\Leftrightarrow B^{-1}\geq A^{-1}\geq 0$.)

As a consequence, market efficiency increases with I. Similarly, for later use, market efficiency decreases if Σ_ε is scaled up with a scalar larger than 1.

Let us finally turn to the macro and micro inefficiencies. The proposition states that

$$\frac{\partial\log(g(n))}{\partial k}>\frac{\partial\log(g(0))}{\partial k}>0. \tag{A.48}$$

Since we already know that $\frac{dI}{dk} < 0$, we are aiming to show that

$$\frac{\partial \log(g(n))}{\partial I^{-2}} > -\frac{\partial \log(g(0))}{\partial I^{-2}}, \tag{A.49}$$

at least for $\frac{\gamma}{I}$ small enough. To that end, we compute $\frac{\partial \log(g(\lambda))}{\partial I^{-2}}$ and check that, at $\frac{\gamma}{I} = 0$, it increases in λ. ∎

Proof of Proposition 7. We concentrate again on interior equilibria. As in the previous proof, the dependence of I and M on k_p is given as a solution to

$$\begin{pmatrix} g_I^M & g_M^M \\ g_I^I & g_M^I \end{pmatrix} \begin{pmatrix} I_{k_p} \\ M_{k_p} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{A.50}$$

and therefore by

$$\begin{pmatrix} I_{k_p} \\ M_{k_p} \end{pmatrix} = \frac{1}{2} \frac{1}{g_M^I g_I^M - g_I^I g_M^M} \begin{pmatrix} g_M^I - g_M^M \\ g_I^M - g_I^I \end{pmatrix}. \tag{A.51}$$

We note that $g_M^I - g_M^M < 0$ and $g_I^M - g_I^I < 0$, while the determinant $g_M^I g_I^M - g_I^I g_M^M$ is negative from (A.44). Thus, both I and M decrease as k_p decreases. Consequently, the inefficiency η increases. Under Assumption $1'$ the macro inefficiency is larger, and more sensitive, than the micro inefficiency.

Since M decreases, while the expression $\frac{M}{S_a} k - c(M, S_a)$ remains equal to zero, S_a must also decrease. The lower cost $f_p = k_p$ of passive investing translates into fewer self-directed investors, leaving an increased number S_p of passive investors. ∎

Proof of Proposition 9. The logic of the proof is the same as for the previous proposition. Specifically, we are solving for the derivatives I_z and M_z from

$$\begin{pmatrix} g_I^M & g_M^M \\ g_I^I & g_M^I \end{pmatrix} \begin{pmatrix} I_z \\ M_z \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \frac{\partial \eta}{\partial z} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{A.52}$$

giving

$$\begin{pmatrix} I_z \\ M_z \end{pmatrix} = \frac{1}{g_M^I g_I^M - g_I^I g_M^M} \begin{pmatrix} g_M^I - g_M^M \\ g_I^M - g_I^I \end{pmatrix} \begin{pmatrix} \frac{1}{2} \frac{\partial \eta}{\partial z} \end{pmatrix}. \tag{A.53}$$

We remarked in the proof of Proposition 6 that η increases with z (holding I fixed). It follows that I and M also increase with z. As in the previous proof, S_a must also increase, while S_p decreases because $S_a + S_p$ does not change. The total effect on the inefficiency η is ambiguous. ∎

Proof of Proposition 8. We need to calculate the utilities u_u and u_i and use the formula

$$\mathrm{E}\left[e^{x^\top Ax+b^\top x}\right] = \det(I_n - 2\Omega A)^{-\frac{1}{2}} e^{\frac{1}{2}b^\top (I_n-2\Omega A)^{-1}\Omega b} \tag{A.54}$$

for $x \sim \mathcal{N}(0,\Omega)$. It helps to actually compute "ex-interim" utilities, conditional on p. Specifically, we compute

$$\max_{x_i} \mathrm{E}\left[e^{-\gamma(x_i(v-p))}\,|p\right] = \mathrm{E}\left[e^{-\frac{1}{2}(\mathrm{E}[v|s]-p)^\top \Sigma_{v|s}^{-1}(\mathrm{E}[v|s]-p)}|p\right] \tag{A.55}$$

by letting $x = \mathrm{E}[v|s] - \mathrm{E}[v|p]$, $A = -\frac{1}{2}\Sigma_{v|s}^{-1}$, and $b^\top = (\mathrm{E}[v|p] - p)^\top \Sigma_{v|s}^{-1}$ to evaluate

$$\begin{aligned}&\mathrm{E}\left[e^{x^\top Ax+b^\top x-\frac{1}{2}(\mathrm{E}[v|p]-p)^\top \Sigma_{v|s}^{-1}(\mathrm{E}[v|p]-p)}\right] \\ &\quad = \det(I_n + \Omega\Sigma_{v|s}^{-1})^{-\frac{1}{2}} e^{\frac{1}{2}(\mathrm{E}[v|p]-p)^\top \Sigma_{v|s}^{-1}(I_n+\Omega\Sigma_{v|s}^{-1})^{-1}\Omega\Sigma_{v|s}^{-1}(\mathrm{E}[v|p]-p)-\frac{1}{2}(\mathrm{E}[v|p]-p)^\top \Sigma_{v|s}^{-1}(\mathrm{E}[v|p]-p)}\end{aligned} \tag{A.56}$$

with $\Omega = \mathrm{Var}\left(\mathrm{E}[v|s]|p\right) = \Sigma_{v|p} - \Sigma_{v|s}$. Simplifying this expression and the analogous one for the uninformed agent shows the equivalence with entropy.

We go further by using the fact that the Kullback-Leibler divergence of a n-dimensional multi-variate normal distribution with mean μ_1 and variance Σ_1 from one with mean μ_0 and variance Σ_0 is

$$D_{KL} = \frac{1}{2}\left(tr\left(\Sigma_1^{-1}\Sigma_0\right) - n + (\mu_1-\mu_0)^\top \Sigma_1^{-1}(\mu_1-\mu_0) + \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\right)\right). \tag{A.57}$$

In our case, $\Sigma_0 = \Sigma_{v|s}$, $\Sigma_1 = \Sigma_{v|p}$, $\mu_0 = \mathrm{E}[v|s]$, and $\mu_1 = \mathrm{E}[v|p]$. Taking expectations, it follows that $\mathrm{E}\left[D_{KL}\right] = \eta$. ∎