Active and Passive Investing: Understanding Samuelson's Dictum

Nicolae Gârleanu and Lasse Heje Pedersen*

Abstract

We model how investors allocate between asset managers, managers choose portfolios of multiple securities, fees are set, and security prices are determined. Investors are indifferent between higher-cost informed managers and lower-cost uninformed managers, interpreted as passive managers as their portfolio is linked to the "expected market portfolio." We make precise Samuelson's dictum by showing that active investors reduce micro-inefficiencies more than they do macro-inefficiencies. 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 the rise of delegated asset management and the resultant changes in financial markets.

JEL: D8, G02, G12, G14, G23, L1

Published version (open access):

Nicolae Gârleanu and Lasse Heje Pedersen, "Active and Passive Investing: Understanding Samuelsons Dictum," *The Review of Asset Pricing Studies*, 2021

https://doi.org/10.1093/rapstu/raab020

*Gârleanu is at Olin Business School, Washington University and NBER and Pedersen is at AQR Capital Management, Copenhagen Business School, and CEPR. We are grateful for helpful comments from Thierry Foucault (editor), Antti Ilmanen, Kelvin Lee, and Peter Norman Sørensen, as well as from seminar participants at the Berkeley-Columbia Meeting in Engineering and Statistics, the Federal Reserve Bank of New York, Copenhagen Business School, the UCLA Anderson Fink Center Conference on Financial Markets, and Georgia State University. 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. Send correspondence to Nicolae Gârleanu, garleanu@wustl.edu.

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 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) active investors reduce micro-inefficiencies more than macro-inefficiencies (consistent with Samuelson's dictum) when there exists a strong common factor in security fundamentals or when the number of securities is large; (3) in fact, all inefficiency is due to systematic factors in the limit with an infinite number of asset, a form of arbitrage pricing theory (APT) for market inefficiency; (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); and (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. 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 and Stiglitz (1980), following Gârleanu and Pedersen (2018). While these models consider

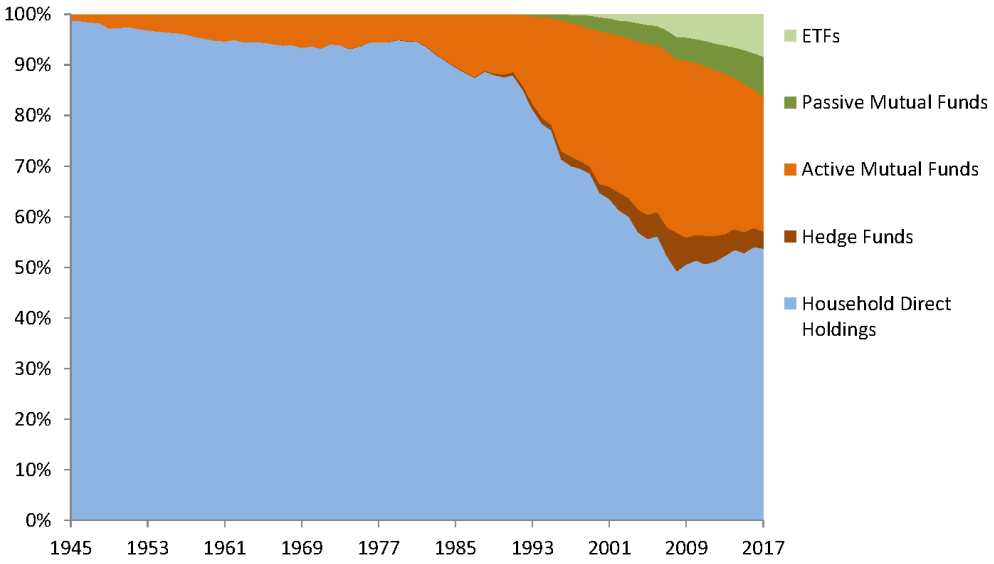

Figure 1: **Ownership of U.S. equities**
The figure shows the decline in direct holdings (blue) and the rise of delegated management, including active mutual funds and hedge funds (the two red areas), passive mutual funds and exchange-traded funds (ETFs) (the green areas), and pension plans and other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

a single risky asset, we consider multiple securities (in the spirit of Admati [1985], who also considers multiple securities, but not asset managers) and, as a further extension of Gârleanu and Pedersen (2018), consider the costs of self-directed investment and passive investment, in addition to the cost of active management.[1] Our core contribution, though, is not the extensions *per se*, but, rather, the novel questions that they enable us to study. Indeed, having asset managers hold multiple securities allows us to study portfolio choice and macro- versus micro-efficiency (neither can be studied with one asset), and having costly active and passive management is essential to study the effects of changes in these costs, for

[1]The model of Gârleanu and Pedersen (2018) features a single risky security and no passive managers and assumes that self-directed trading is costless. See also García and Vanden (2009), who consider 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).

example, as fintech reduces the costs of index funds and ETFs.

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 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, noise from hedging demand, 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 passive managers 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 usually 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)). Bridging the REE literature and the CAPM, Admati (1985) points out that the unconditional expected market portfolio is generally not the optimal portfolio for uninformed investors. Indeed, uninformed investors can do better by using the information reflected in prices, as shown theoretically and empirically by Biais et al. (2010). Extending these insights, we consider whether uninformed investors optimally hold the closest thing that they can get to the market portfolio, namely, the "conditional expected market portfolio" based on the distribution of the supply and what can be learned from prices (and other public signals). The answer is "yes" under certain special conditions (e.g., i.i.d. shocks across securities), which resembles what real-world passive investors do, namely, choosing an index that is rebalanced based on public information. However, indexes 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 uninformed investors optimally overweight securities with less supply uncertainty in the more general case in which shocks are correlated across securities via a factor structure. Hence, our framework presents a step toward a theory of optimal security indexes.

Active investors 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, market inefficiency is zero (fully efficient market) if uncertainty about the fundamental value is the same whether one learns only from the price or also from the signal. Further, the larger information advantage one enjoys from knowing the signal, the more inefficient the market.

An interesting question naturally addressed in our framework is whether active investors make the market more efficient at the micro level than at the macro level, as Samuelson famously hypothesized (see quote and references in the beginning of Section 2 and the evidence in Jung and Shiller (2005) and Xiao et al. (2021)). The idea, known as "Samuelson's dictum," is that active investors have stronger incentives to correct (micro) inefficiencies in relative prices than to correct the overall (macro) price level. For example, active investors will ensure that the price difference between General Motors and Ford is close to efficient, but active investors may leave all stocks overvalued.

We show that Samuelson's dictum holds when a strong common factor exists in the 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, because of diversification, when the number of securities is large, Samuelson's dictum actually always holds.

In fact, any micro-inefficiency goes to zero as the number of assets grows. Perhaps

surprisingly, the *combined* inefficiency of all micro portfolios becomes negligible with many assets. Almost all the inefficiency is in the pricing of systematic factors. This result is related to the arbitrage pricing theory (APT) of Ross (1976). While APT states that risk premiums are driven by systematic factors when the number of assets is large, we show that inefficiencies are also driven by these factors.

The key simplifying assumption that allows these striking results is that active managers in our model make an all-or-nothing information choice, which captures the idea that active investors must decide whether or not to set up an IT system that captures the main databases and a staff that can process all these data. In contrast, Veldkamp (2011), Van Nieuwerburgh and Veldkamp (2010), Kacperczyk et al. (2016), and Glasserman and Mamaysky (2018) study agents' choice of information, which can affect macro- versus micro-efficiency as emphasized by the latter paper.[2] We complement the literature with regard to macro- versus 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 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.

Samuelson also hypothesized that efficiency, especially micro-efficiency, has improved over time (see quote and reference in Section 3). 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 consistent with the empirical findings of Bai et al. (2016), Dávila and Parlatore (2018), 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

[2]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.

of self-directed investment and active management), consistent with the development in the 2000s.[3] Further, a 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 lead to a larger share of passive investment, a higher average alpha for active managers, and lower fees for active investing, but an increased fee spread between active and passive managers.

In summary, we make precise Samuelson's dictum and show that it holds when the number of assets is large enough, illustrating that our REE model with a large number of assets is a powerful tool. 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.[4]

1 Model and Equilibrium

This section lays out our model and shows how to solve it.

1.1 REE model with multiple assets and asset managers

Markets. 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

[3] In a similar spirit, Peress (2005) shows that a decreasing market-participation cost leads to more participation and, in particular, more passive participation.

[4] 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 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).

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.[5] We do not think of the signal as (illegal) insider information, but rather as the ability to access and process large amounts of relevant data, which could be available for purchase or even freely. In other words, informed investors need to acquire databases, collect other relevant information, and set up systems to optimize their portfolio in light of these signals. This is a costly process, as securities databases can be expensive and trading continually on all this information requires an investment staff. We note that, as in Grossman and Stiglitz (1980), investors must simply decide whether or not to acquire this information, which captures the idea that information processing can be seen as an infrastructure with a fixed cost while the papers cited in the introduction consider endogenous information precision or choice.

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

[5] 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 $\Sigma_u = \mathrm{var}(u)$.

[6] Many indexes use a float adjustment. For example, S&P Float Adjustment Methodology 2017 states "the share counts used in calculating the indexes 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."

Investors. The economy has $\bar{S}$ investors with initial wealth W and constant absolute risk aversion (CARA) coefficient γ. These investors search for an active manager, allocate to a passive manager, or directly invest in the financial market. Specifically, I investors choose to search for an informed active manager, S_p investors choose passive management, and the remaining $\bar{S} - I - S_p$ 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 a cost d_l associated with brokerage fees and time used on portfolio construction, and these costs vary across investors. If the investor makes an informed investment, the cost is $d_l + k$, but we show that, for what we argue to be reasonable parameters, this behavior is dominated by using an active manager due to the effective ability to share information costs across the manager's investment base. Therefore, the total number of uninformed investors, U, is the sum of the number of self-directed investors and investors with passive managers, that is, everyone who is not searching for an informed manager, $U = \bar{S} - I$.

If the investor uses a passive manager, he incurs the passive management fee f_p. Finally, if the investor chooses to search for an informed active manager, the investor must pay a search cost c and, in addition, pay an active asset management fee f_a determined via Nash bargaining. The search cost $c(M, I)$ is a smooth function satisfying $\frac{\partial c}{\partial M} \leq 0$ and $\frac{\partial c}{\partial I} \geq 0$, meaning finding an appropriate manager is easier to do if there are more managers and more difficult when more agents are searching for a manager.[7]

To understand why investors may choose to use asset managers, note first that an investor with a higher cost (d_l) of self-directed investing than the passive manager's marginal cost (k_p) clearly benefits from outsourcing any passive portfolio construction to the manager. Further, investors with active managers can benefit from both potential low marginal costs from the manager (k_a) and, importantly, the fact that the active manager can effectively share the fixed cost of information (k) across all his investors. On the other hand, an in-

[7]An analytically convenient search function that satisfies these requirements is $c(M, I) = \bar{c}\left(\frac{I}{M}\right)^{\alpha}$ for constants $\bar{c} > 0$ and $\alpha > 0$. Investors' search cost capture expenses of vetting a manager. Gârleanu and Pedersen (2018) describe in their appendix B how real-world investors search and perform due diligence. While searching ensures finding an informed manager in our model, real-world search is imperfect.

vestor with very low cost of self-directed investing may be best off investing himself.

Asset managers. 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").

Portfolios and utilities. Informed managers choose a portfolio that is optimal for their investors conditional on the signal, denoted x_i where i stands for informed.[8] The other investors (i.e., the self-directed and those investing with a passive manager) acquire a portfolio that is optimal conditional on observing only the price, denoted x_u where u means uninformed. The resultant certainty equivalent utility (before fees and other costs) for an investor is

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

where $\mathcal{F}_j$ is the information set used in portfolio selection with $j \in \{i, u\}$, that is, informed/uninformed. The above equality defines the certainty equivalent utilities of being informed, u_i, and uninformed, u_u, as well as the corresponding optimal portfolios x_i and x_u.

Equilibrium. 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 investors who allocate to

[8] In our model, incurring a search cost alleviates any agency issues. For a recent model of agency issues in asset management, see Buffa et al. (2020).

active I or passive managers S_p such that (a) the supply of shares equals the demand

$$q = Ix_i + Ux_u, \tag{2}$$

(b) fees are set via Nash bargaining; (c) each active manager decides optimally whether to be informed; and (d) each investor decides optimally whether to use an active manager, use a passive manager, or be self-directed.

We show below how to solve the model.[9] We conjecture and verify an equilibrium in which asset prices p are linear in the information s about securities as well as the supply q:

$$p = \theta_0 + \theta_s \left((s - \bar{v}) - \theta_q (q - \bar{q}) \right), \tag{3}$$

where θ_0, θ_s, and θ_q are real parameters to be determined.

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), who consider 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 $l = 1$, that is, 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.

$$\eta^{\zeta} = \frac{1}{2} \log \left(\frac{\det(\text{var}\left(\zeta^\top v | \mathcal{F}_u\right))}{\det(\text{var}\left(\zeta^\top v | \mathcal{F}_i\right))} \right), \tag{4}$$

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

[9] In general, an equilibrium with nonzero I, 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.

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

$$\eta^{\text{asset 1}} = \frac{1}{2} \log \left(\frac{\text{var}\left(v_1 | \mathcal{F}_u\right)}{\text{var}\left(v_1 | \mathcal{F}_i\right)} \right) = \log \left(\frac{\text{var}\left(v_1 | \mathcal{F}_u\right)^{1/2}}{\text{var}\left(v_1 | \mathcal{F}_i\right)^{1/2}} \right). \tag{5}$$

We see that inefficiency is zero if the market is perfectly efficient in the sense that the price reveals all information about the signal. Otherwise, the inefficiency is a positive number. A lower inefficiency corresponds to a higher price informativeness as defined by Grossman and Stiglitz (1980). Indeed, Grossman and Stiglitz (1980) define the price as more informative if it is more correlated with the signal since, in this case, uninformed agents who observe the price can learn more about the signal. This higher informativeness further implies that informed and uninformed agents have more similar information and therefore similar uncertainty, meaning that market inefficiency is lower by our definition (5).

When we analyze macro- versus micro-efficiency (in Section 2), we also study the efficiency of individual securities, portfolios, and collections of portfolios, so we need a general definition of efficiency. The correlation-based definition of Grossman and Stiglitz (1980) does not lend itself to be generalized to many assets and portfolios of assets, but our definition is both general and equivalent to Grossman and Stiglitz (1980) in the one-asset case. Further, Grossman and Stiglitz (1980) show that informativeness is linked to the ratio of the utilities of the informed and uninformed agents, and our Proposition 1 shows that our more general definition of inefficiency retains this idea.[10]

[10]Grossman and Stiglitz (1980) define the informativeness of the price of a single asset as the squared correlation between the price and the signal in their theorem 4.A (among other places). Combining their equations (13), (17), and (18) shows that this informativeness is a decreasing function of the ratio of the utilities of the informed and uninformed agents, which in turn is equivalent to our definition of inefficiency, as seen from our Proposition 1. The Grossman and Stiglitz (1980) informativeness is thus essentially the inverse of our inefficiency. The link between the value of information and the ratio of determinants of the conditional variances was first derived in Admati and Pfleiderer (1987).

Overall market inefficiency plays a special role in the equilibrium of the model. Overall market inefficiency is naturally the inefficiency of the set of all assets. Hence, we consider the largest possible matrix of portfolios, $\zeta = \mathrm{I}_n$, namely, the identity matrix in $\mathbb{R}^{n\times n}$.[11] We denote overall market inefficiency simply by η:

$$\eta = \eta^{\text{overall market}} = \eta^{\mathrm{I}_n} = \frac{1}{2}\log\left(\frac{\det(\text{var}\,(v|\mathcal{F}_u))}{\det(\text{var}\,(v|\mathcal{F}_i))}\right). \tag{6}$$

This definition of overall market inefficiency is the natural extension of the one-asset definition of Grossman and Stiglitz (1980), since it maintains the tight link between market inefficiency and investors' utility of information as discussed further below and formalized in Proposition 1.

Proposition 1 (Efficiency and the value of information) *The utility difference between informed and uninformed investors equals overall market inefficiency,*

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

Furthermore, for any set of portfolios $\zeta \in \mathbb{R}^{n\times l}$, if an informed and uninformed investor were restricted (out of equilibrium) to trading only ζ, then the utility difference would be $\gamma(u_i - u_u) = \eta^{\zeta}$.

1.3 Solution: Deriving the equilibrium

Active versus passive investing. We start by considering the optimal behavior of each investor indexed by l, recalling that investors differ in their cost, d_l, of direct investment. An investor l with a cost of direct investment that is lower than the passive fee, $d_l < f_p$, optimally invest directly in the financial market. Investors with higher costs of direct investment, $d_l \geq f_p$, find it suboptimal to invest directly and, in an interior equilibrium, are indifferent between active and passive management. The indifference condition for these investors equalizes the certainty equivalent utility of passive management ($W + u_u - f_p$) with that of active

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

management ($W + u_i - c - f_a$),

$$u_i - u_u = f_a + c - f_p \tag{8}$$

This equilibrium condition 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 is the certainty equivalent utility of getting an informed portfolio rather than an uninformed one. The net cost of active investing is the active fee plus the search cost minus the passive fee. Next, we will show how to compute these costs and benefits as functions of the number of informed investors, I. Figure 2 illustrates how the costs and benefits of active investing depends on the number of active investors. When these costs and benefits are computed, we can find an interior equilibrium as the intersection of the lines in Figure 2.

The benefit of active investing: Market inefficiency. We first consider how the benefits of active investing depends on the number of active investors, that is, the right-hand side of (8). The utility benefit of being informed relative to being uninformed as already characterized in Proposition 1. This utility difference equals η/γ, that is, overall market inefficiency divided by risk aversion, to measure it in certainty-equivalent value. We show in the appendix how to compute the market inefficiency along with the equilibrium price function (3):

$$\eta = \frac{1}{2}\log\left(\frac{\det\left(\Sigma_v^{-1} + \Sigma_\varepsilon^{-1}\right)}{\det\left(\Sigma_v^{-1} + (\Sigma_\varepsilon + \frac{\gamma^2}{I^2}\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1}\right)}\right). \tag{9}$$

This equation shows explicitly how the market inefficiency depends on the number of active investors I (as well as the exogenous variance-covariance matrices). We see from (9) that market inefficiency decreases in I as is also clear from Figure 2. This finding is intuitive since more active investors lead to more informative prices, and, hence, lower market inefficiency.

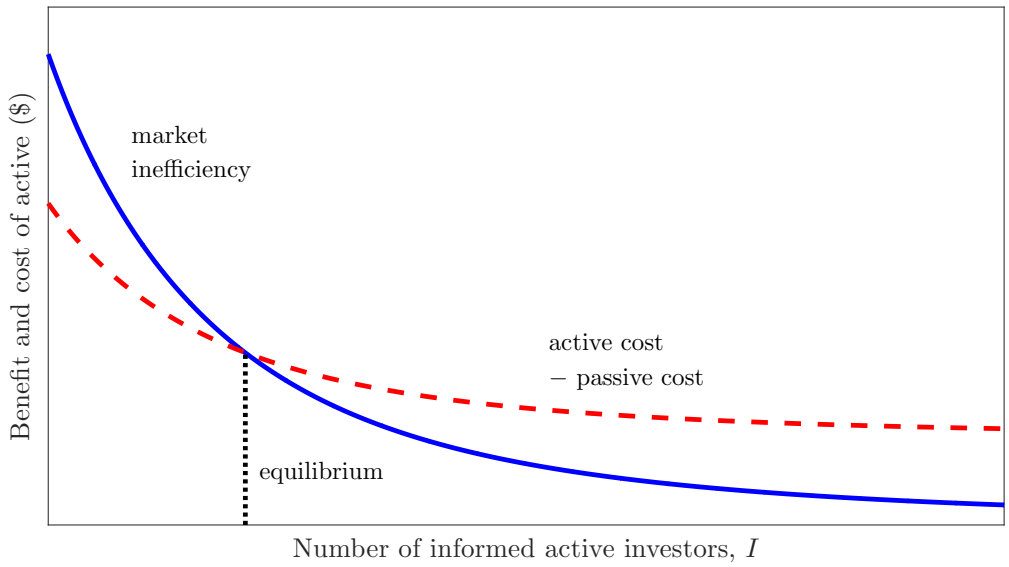

Figure 2: **Equilibrium market inefficiency and number of active investors**
The figure shows how an equilibrium is found by equalizing the cost and benefit of active investing. The benefit of active investing (the solid line) stems from the ability to exploit market inefficiency; expressed as certainty equivalent wealth, it equals $u_i - u_u = \eta/\gamma$. The net cost of active investing (the dashed line) equals the active fee plus the search cost, minus the cost of passive investing.

The cost of active investing. Next, we compute the cost of active investing relative to the cost of passive investing, that is, the right-hand side of Equation (8). We first recall that the passive management fee equals the marginal passive cost $f_p = k_p$.

Next, we turn to the active-management fee, f_a. 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 appears in both terms because it is an already sunk cost.[12] The manager's gain from accepting the investor is her fee revenue f_a minus his

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

marginal cost k_a. Hence, the equilibrium fee is

$$\begin{aligned} f_a &= \arg\max_f \left(W - c - f + u_i - (W - c - f_p + u_u)\right)(f - k_a) \\ &= \arg\max_f \left(u_i - u_u + f_p - f\right)(f - k_a) = \frac{u_i - u_u + f_p + k_a}{2} \\ &= \frac{k_a + k_p}{2} + \frac{\eta}{2\gamma}, \end{aligned} \tag{10}$$

where we use (7). 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 final cost of active investing is the search cost, $c(M, I)$. The search cost depends on the number of informed managers M and the number of active investors I, so we need to consider how many managers become informed in equilibrium.

A manager that acquires the signal and becomes an informed active manager expects to attract I/M active investors. These investors give rise to an income, net of marginal costs, of $(f_a - k_a)I/M$. In contrast, an uninformed manager expects no investors. Hence, the active manager's indifference condition for paying the information cost k is $k = (f_a - k_a)I/M$, which we rewrite using (10) as

$$M = \frac{I}{k}\left(\frac{\eta}{2\gamma} + \frac{k_p - k_a}{2}\right). \tag{11}$$

In summary, we see that the cost of active investing depends on I because (a) the active fee depends η as seen in (10), which in turn depends on I as seen in (9); and (b) the search cost depends on I and M, where the latter depends on I directly and via η as seen in (11). The cost of active investing is illustrated in Figure 2, where it is seen to decrease in the number of active investors. The reason is that more active investors means more efficient markets, leading to lower active fees. The cost of active fees can also increase in I, though, if the search cost rises sufficiently.

Computing the equilibrium. Having explicitly computed the cost and benefit of active investing as a function of I, we can find the equilibrium value of I by equating the cost and benefit, a nonlinear algebraic equation that is easy to solve numerically for any search-cost function $c(M, I)$.[13]

Proposition 2 *In an interior equilibrium, the number of active investors, I, is a solution to*

$$\frac{\eta(I)}{2\gamma} = \frac{k_a - k_p}{2} + c\left(\frac{I}{k}\left(\frac{\eta(I)}{2\gamma} + \frac{k_p - k_a}{2}\right), I\right) \tag{12}$$

where the market inefficiency $\eta(I)$ is given by (9)*. The corresponding active asset management fee f_a is given by* (10)*, the number of informed active managers M is seen in* (11)*, and the price function is given in Appendix A.3.*

1.4 Quantitative implications

Equation (10) provides a simple relation between market inefficiency, on the one hand, and active and passive management fees, on the other. Here we exploit this relation to make quantitative statements about the former based on observations of the latter.

We combine Equation (10) and the equation for the passive fee, $f_p = k_p$, to express overall market inefficiency, η, as

$$\eta = 2\gamma(f_a - f_p) + \gamma(k_p - k_a) = 2\gamma^R({f_a}^{\%} - {f_p}^{\%}) + \gamma^R({k_p}^{\%} - {k_a}^{\%}), \tag{13}$$

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

[13] Such a solution can be guaranteed to exist by either increasing $\bar{M}$ or scaling down the cost function sufficiently. We assume hereafter the interesting case in which an interior equilibrium exists.

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 and market risk premium, both of which are endogenous. Just as the CAPM provides useful insights on expected returns, our calculation provides an interesting implication on the magnitude of market efficiency based on the level of fees that investors are willing to pay.

The model is, of course, richer, and provides both the fees and the market inefficiency as outcomes once the asset structure is fully specified and parameterized, which we do starting in the next section.

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 (13), we predict that overall market inefficiency changes by

$$\begin{aligned} \Delta\eta &= 2\gamma(\Delta f_a - \Delta f_p) + \gamma(\Delta k_p - \Delta k_a) \\ &= 6 \times (-0.16\% - (-0.50\%)) + 3 \times (-0.50\% - 0) = 0.54\%. \end{aligned} \tag{14}$$

Naturally, cheaper index funds increases passive investing and decreases active investing, leading to higher equilibrium market inefficiency as seen in (14). This increase in market inefficiency leads to higher expected returns to active management, consistent with the findings of Cremers et al. (2016). The intuition behind this finding is seen in Figure 2, where the reduction in passive costs corresponds to an upward shift in the dashed line. Further, the model can help quantify the magnitude of the effect as seen in (14). We will illustrate these effects further with numerical examples in Sections 2 and 3.

1.5 Understanding passive investment

In standard theories of informed trading, investors naturally fall into two groups—informed and uninformed—while real-world investors are classified as active or passive, and we have made the natural link between these dichotomies. The connection between "informed" and "active" investors seems uncontroversial and their portfolios are described further in Appendix A.1.

The link between "uninformed" and "passive" investors deserves more consideration. What these concepts clearly have in common is that the uninformed investors of abstract models minimize their costs as do real-world passive investors. Uninformed investors seek to maximize their performance subject to a minimal cost, which is presumably also the goal of passive investors. But how can we make this connection clearer?

First, note that while passive indexes typically follow relatively simple rules, competition between index providers could lead these rules to become optimal subject to minimum costs, just like the portfolios of uninformed investors in the model.

Second, passive investors are doing something slightly more complicated than "just buying the market" (as would be the prediction of the CAPM), because, in the real world, no one knows exactly what the market portfolio is (Roll, 1977) and it is constantly changing. In particular, passive investors must choose how to be passive, for example, choosing a portfolio of (global) stock and bond index funds or ETFs. These indexes are typically weighted by the total market capitalization of each constituent security's floating shares outstanding, and they are regularly reweighted 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).

Likewise, uninformed investors do not know the market portfolio q in our noisy REE economy. So, how do uninformed investors choose their portfolio, x_u, defined in Equation (1)? Clearly, x_u only depends on prices and public information, such as the distribution of shares outstanding, so could there be link to real-world indexes? In this connection, it is interesting to relate the uninformed portfolio x_u to the uninformed investors' best estimate of the true

market portfolio, q, based on public information. We call this best estimate the *"conditional expected market portfolio,"* $E(q|p)$.

So uninformed and passive investors both minimize cost, presumably both maximize their performance, and perhaps both have portfolios based on $E(q|p)$, which we explore based on following assumptions.

Assumption 1 *Fundamentals have a factor structure:*

$$v = \bar{v} + \beta F_v + w_v, \tag{15}$$

$$\varepsilon = \beta F_\varepsilon + w_\varepsilon, \tag{16}$$

$$q = \bar{q} + \beta F_q + w_q, \tag{17}$$

where $\bar{v}, \bar{q} \in \mathbb{R}^n$ *are the average fundamental values and supplies, respectively,* $\beta \in \mathbb{R}^n$ *is a vector of factor loadings, 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* $\sigma^2_{w_v}$*,* $\sigma^2_{w_\varepsilon}$*, and* $\sigma^2_{w_q}$*, respectively, for each asset.*

Assumption 1′ *Assumption 1 holds,* $\sigma^2_{F_v} > 0$*, and* $\sigma^2_{F_v}/\sigma^2_{w_v} \geq \sigma^2_{F_\varepsilon}/\sigma^2_{w_\varepsilon}$*.*

Assumption 1 means that the model variables are driven by a standard factor model. We 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 unconditional average market portfolio, $\bar{q}$. The factor portfolio also could be different, but we normalize β so that $\beta^\top \bar{q} \geq 0$. Assumption 1′ means that the common factor-component is especially important for future fundamentals, v. We note that, for simplicity, the factor β is the same for the different types of shocks in (15)–(17), but we later consider a more general version with different betas in Assumption 3. 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.[14] 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.

Proposition 3 (Optimal uninformed portfolio) *(a) The optimal uninformed portfolio, x_u, is proportional to the conditional expected market portfolio, $E(q|p)$, in that there exists $A \in \mathbb{R}^{n\times n}$ such that*

$$x_u = A\,E(q|p). \tag{18}$$

Under Assumption 2, $A \in \mathbb{R}_+$ is a scalar. Under Assumption 1, A is positive definite and

$$x_u = A_0\,E(q|p) - A_1\,\beta\beta^\top E(q|p). \tag{19}$$

where $A_0 \in \mathbb{R}_+$ and A_1 is a scalar, which is also positive if Assumption 1′ holds.

(b) The average uninformed portfolio, $E(x_u)$, is proportional to the average market portfolio, $\bar{q}$, in that $\bar{A} \in \mathbb{R}_+$ exists such that $E(x_u) = \bar{A}\,\bar{q}$ under Assumption 2 or if Assumption 1 holds and β and $\bar{q}$ are collinear. Alternatively, 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

[14] To understand this assumption, consider what happens if any security j undergoes a two-for-one 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.).

safer asset will have a larger weight in the average passive portfolio, $E(x_{u,i}) \leq E(x_{u,j})$*. As another alternative, 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.*

We see that the optimal uninformed portfolio can indeed be linked to the conditional expected market portfolio, $E(q|p)$. As seen in part 1 of Proposition 3, uninformed investors use public information to proxy for the expected market portfolio as an input to the portfolio construction, similar to real-world indexes. Further, part 1 states that, under certain conditions, the optimal passive portfolio is literally just that, namely, the conditional expected market portfolio. However, 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 further shows that, even when the uninformed portfolio is not equal to the expected market portfolio state by state, portfolios may coincide on average. In the more general situation in which β and $\bar{q}$ are not proportional, the optimal passive portfolio tends to downweight risky securities or securities with more supply uncertainty. This intuition may help explain why real-world passive indexes only include a subset of listed securities, which may be interpreted as a binary version the results of part 2. 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, indexes typically exclude securities with too low price, too low market capitalization, too low liquidity, too recent initial public offering, or involved in certain corporate actions.[15]

[15]See, for example, and "Russell U.S. Equity Indexes 2017" and "S&P U.S. Indices Methodology 2017."

2 Samuelson's Dictum: Macro- versus Micro-Efficiency

If inefficiencies are the "crimes" in financial markets, active investors are the "police" who correct these offenses. Therefore, the overall level of market efficiency is linked to the costs and fees of active asset management, relative to the cost of passive management, as seen in Equation (10). But which securities and portfolios are made especially efficient by active investors, and which are left relatively inefficient? In other words, what "crimes" are active investors best at fighting?

From the perspective of uninformed passive investors, how well are these investors "protected" from buying a security or portfolio at an inefficient price? Do passive investors face a greater risk of buying all stocks at an inefficient overall price level, or a greater risk of buying General Motors for too much relative to the price of Ford?

In this connection, Paul Samuelson famously conjectured that markets would have greater micro-efficiency than macro-efficiency:[16]

> 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 (4).

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

[16]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- versus micro-efficiency appear in, for example, Samuelson (1998).

inefficiency in equilibrium (i.e., after active investors have done their trading):

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

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.

2.1 Principal inefficiency portfolios

To solve the problem of maximizing inefficiency, we let $G = \text{var}(v|s)^{-1/2}\text{var}(v|p)\text{var}(v|s)^{-1/2}$. This matrix G captures 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 information-advantage 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 call $\hat{w}_1$ the first principal inefficiency portfolio.

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 overall market inefficiency in terms of the eigenvalues (g_j):

$$\eta = \frac{1}{2}\log\left(\frac{\det(\text{var}(v|p))}{\det(\text{var}(v|s))}\right) = \frac{1}{2}\log\left(\det(G)\right) = \frac{1}{2}\sum_{j=1}^{n}\log g_j = \sum_{j=1}^{n}\eta^j, \tag{20}$$

where η^j is the inefficiency of portfolio $\hat{w}_j$, defined analogously to $\hat{w}_1$, as the principal inefficiency portfolios, $\hat{w}_j = \text{var}\left(v|s\right)^{-1/2} w_j$. We see that overall market inefficiency is the sum of portfolio inefficiencies for the set $(\hat{w}_j)$ of independent[17] portfolios that spans the space of all portfolios.

To understand the economics of Equation (20), 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 $\frac{1}{2}\log g_1$.[18] 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 $\eta = \frac{1}{2}\sum_j \log g_j$.

2.2 Samuelson's dictum: A finite number of assets

We are ready to state our first result on macro- versus micro-efficiency.

Proposition 4 (Macro- versus Micro-Efficiency)

(a) **(Samuelson's dictum)** *Under Assumption 1′, the most inefficient portfolio is the*

[17] While the original eigenvectors (w_j) are orthogonal in the Euclidean norm, $w_j^\top w_k = 0$, the corresponding inefficiency portfolios $(\hat{w}_j)$ are orthogonal in the economically more interesting sense that their payoffs are conditionally uncorrelated, $\text{cov}(\hat{w}_j^\top v^\top, \hat{w}_k^\top v|s) = \hat{w}_j^\top \text{var}\left(v|s\right)\hat{w}_k = 0$. (We also note that, under Assumption 1, the rotation only scales the eigenvectors w_1 through w_n and the corresponding portfolios remain orthogonal in the Euclidian norm.)

[18] 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, $\frac{1}{2}\log g_1$.

factor portfolio,

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

and the least inefficient portfolios are those that eliminate factor risk, that is,

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

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

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

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

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

Part (a) 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. The most inefficient portfolio is the factor portfolio, that is, the portfolio with the most systematic risk. The least inefficient portfolios are long-short portfolios that eliminate all factor risk. This result arises because active investors who "spot aberrations from micro-efficiency can make money from those occurrences and, in doing so, they tend to wipe out any persistent inefficiencies." Indeed, Samuelson's words describe well what goes on in our model. An active 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. In contrast, active investors leave the factor portfolio relatively inefficient because, while they can potentially learn a lot about its fundamental value (since signal noise cancels out to some extent), trading on this information is risky because the factor risk is systematic. 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.[19]

Further, the proposition provides conditions under which Samuelson's dictum applies. The sufficient condition 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. While this condition appears empirically plausible, we believe that an even stronger argument for Samuelson's dictum is that it always holds when the number of securities is large, which would certainly be an empirically relevant condition. We provide a formal statement in the next section.

Part (b) of the proposition states that, under certain conditions, all portfolios are equally efficient. This conclusion applies, for example, when all securities are independent (in terms of the fundamental values, signals, and supply shocks). This is perhaps not so surprising since, with independent securities, it is difficult to even define micro and macro, but this conclusion even holds in cases when micro and macro can be defined under Assumption 2 (although, as explained above, Assumption 2 is unrealistic for other reasons).

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 $\sigma^2_{F_v}$, the variance $\sigma^2_{F_\varepsilon}$ of the common component in the signal noise is sufficiently large.[20]

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- versus 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.

In summary, we have seen how to make Samuelson's dictum precise as the statement

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

[20] In this case, the correlated signals convey little information about the factor portfolio. Indeed, at the limit, as $\sigma^2_{F_\varepsilon}$ becomes infinite, the inefficiency is zero: no information is 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.

that active investors make factor-neutral portfolios most efficient, while leaving the factor portfolio as the most inefficient portfolio. We have also seen that the dictum holds only under certain conditions, but we will next show that the dictum always holds when the number of securities is large enough.

2.3 Samuelson's Dictum meets arbitrage pricing theory

Next, we consider what happens to macro- and micro-efficiency when the number of assets, n, is large. Indeed, thousands of securities exist in the real world, so we may achieve a tractable approximation of the real world by considering the simplifications that arise when we let n go to infinity. Further, the tractability afforded by looking at this large-asset limit also allows us to consider a more general factor structure with multiple factors, similar to the setting of the Ross (1976) arbitrage pricing theory (APT). Hence, we can study the pricing of risk and market efficiency when active investors can diversify across idiosyncratic risks.

So far, we have relied on Assumption 1 with a single factor, which is the same for payoffs v, noise ε, and supply q. We now consider Assumption 3, which allows multiple factors that differ across v, ε, and q. In the interest of readability, we state the assumption without making explicit the dependence of variables on n (and we sacrifice some of the possible generality).

Assumption 3 *For any n, the payoff v, noise ε, and supply $q = q'/n$ are*

$$v = \bar{v} + \beta_v F_v + w_v \tag{21}$$

$$\varepsilon = \beta_\varepsilon F_\varepsilon + w_\varepsilon \tag{22}$$

$$q' = \bar{q}' + \beta_{q'} F_q + w_{q'} \tag{23}$$

where $\bar{v}, \bar{q}' \in \mathbb{R}^n$, $\beta_v \in \mathbb{R}^{n \times k_v}$, $\beta_\varepsilon \in \mathbb{R}^{n \times k_\varepsilon}$, and $\beta_{q'} \in \mathbb{R}^{n \times k_q}$, and the columns of β_ε linearly generate those of β_v. The common factors $F_v \in \mathbb{R}^{k_v}$, $F_\varepsilon \in \mathbb{R}^{k_\varepsilon}$, and $F_q \in \mathbb{R}^{k_q}$ are independent multivariate standard normal variables. The idiosyncratic shocks w_v, w_ε, and $w_{q'}$ are independent normal random variables with zero means and variances that are bounded above,

and below away from zero, uniformly in n. The following limits are well defined and finite: $\lim_{n\to\infty} n^{-1}\bar{v}^\top \bar{q}'$, $\lim_{n\to\infty} n^{-1}\beta_v^\top \bar{q}'$, *and, for any x and y in $\{v, \varepsilon, q'\}$,*

$$\lim_{n\to\infty} n^{-1}\beta_x^\top \beta_y = B_{xy} \in \mathbb{R}^{k_x \times k_y}, \tag{24}$$

with B_{vv} and $B_{\varepsilon\varepsilon}$ strictly positive definite matrices and $B_{q'v} \neq 0$.

To understand Assumption 3, note that (21)–(23) let factor loadings differ across payoffs v, noise ε, and supply q, and (24) is a regularity condition on the factor loadings to ensure convergence as n grows.[21] We write the supply $q = q'/n$ in terms of an auxiliary variable q' for notational ease. In particular, q'_i is of similar size and variance across securities i, but the economy would not converge if we had a fixed group of investors who had to own more and more securities as n grows (since the risk per investor would go to infinity). Therefore, as we increase the number n of securities, we simultaneously reduce the supply per security by letting $q = q'/n$.[22]

As another motivation of Assumption 3, the next proposition shows that it implies a version of the standard APT. The standard APT describes expected returns in the presence of a factor structure, and we show that our factor structure is scaled in a way that is consistent with an APT equilibrium with many assets. The more novel part of the proposition is a related APT result for market inefficiency. To understand the APT for efficiency, recall that the inefficiency η^ζ of any set of portfolios ζ is always less than or equal to overall market inefficiency, η.

Proposition 5 *Under Assumption 3, the following statements hold.*

[**APT of Returns**] *The market risk premiums are determined by fundamental factor load-*

[21] The assumption that the columns of β_ε linearly generate those of β_v excludes the case in which a well-diversified portfolio can be constructed with positive systemic fundamental risk, but no systemic signal noise. The conditions on the matrices B_{xy} exclude degenerate cases, and in particular ensure that the limit market inefficiency is not zero.

[22] One special case of Assumption 3 is the following: Start with an economy with a finite number of assets with a factor structure, and then make an increasing number of "copies" of all stocks in this baseline economy. When there are m copies of each baseline stock, the supply of each stock is $1/m$ so that the total payoff in the economy is roughly unchanged with the number of copies.

ings, that is, there exist $\lambda_1, ..., \lambda_{k_v} \in \mathbb{R}$ *so that* $E(v_i - p_i) = \sum_{j=1}^{k} \beta_{v,ij}\lambda_j$ *up to an error term with* ℓ^2 *norm that is uniformly bounded in* n*. If, furthermore,* $\|\bar{q}\|_2 \to_n 0$*, then these errors term tend to zero, that is, beta pricing is exact. Consequently, a portfolio with zero loadings on all* F_v *factors has zero expected excess return.*

[**APT of Efficiency**] *The fraction of the market inefficiency coming from the systematic fundamental factors approaches 100% as* $n \to \infty$*, that is,* $\eta^{\beta_v}/\eta \to 1$*. In contrast, any portfolio with zero loadings on all* F_v *factors has zero inefficiency in the limit.*

The standard "APT of returns" says that risk premiums 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. As in Ross (1976), the APT generally holds as an approximation, but having a fully specified equilibrium provides us with an explicit condition for it to hold exactly, namely, that every asset constitutes a vanishing part of the economy in the limit.

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.

To fully appreciate the result in Proposition 5, note that the overall inefficiency η can be seen as the sum of the inefficiencies of n uncorrelated portfolios as given by Equation (20). With the assumed factor structure, the overall inefficiency η equals the inefficiency of the factor portfolio η^{β_v} plus the sum of inefficiencies of $n - k_v$ uncorrelated micro portfolios.[23] Our result is therefore not simply that each individual micro-inefficiency goes to zero; rather, the *sum* of all micro-inefficiencies goes to zero, even though the number of micro portfolios goes to infinity. In other words, factor inefficiency dominates overall market inefficiency to a surprising extent.

[23] If the factor portfolios given by the columns of β_v are orthogonal, then these portfolios are in fact the same as the first principal inefficiency portfolios described in Section 2.1, that is, $\eta^{\beta_{v,j}} = \eta^j$. In this case, we can write the APT of efficiency as $\sum_{j=1}^{k} \eta^{\beta_j}/\eta \to 1$, meaning that the sum of inefficiencies of factor portfolios is nearly the entire market inefficiency.

The fact that micro portfolios, which load exclusively on idiosyncratic shocks, are asymptotically efficient is not difficult 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 the combined inefficiency of all micro portfolios tends to zero is more involved and more surprising since the number of micro portfolios goes to infinity as n increases. The reason for this result is that a hypothetical active investor who is restricted to trading micro portfolios has a utility gain over that of an uninformed investor that declines as $1/n$ and thus goes to zero as n increases.[24] On the other hand, when trading fundamental factors, an informed investor has a utility gain relative to an uninformed that is bounded away from zero since the supply noise in these factors does not vanish as n goes to infinity.

The proposition states that a number of portfolios equal to that of fundamental factors (k_v) is sufficient to exploit the entire inefficiency in the limit. There are infinitely many such sets of portfolios. With more structure than imposed by Assumption 3, statements can be made about the maximal-inefficiency portfolios for finite and large n. In particular, in some cases the maximal efficiency sets of portfolios for any finite n sufficiently large can be identified explicitly. We illustrate this claim with some simple examples.

Example 1: Single-factor model. Consider a one-factor market satisfying Assumption 3 with $\beta_v = \sigma_{F_v} 1_n$, where $1_n := (1, 1, ..., 1)^\top \in \mathbb{R}^n$. Similarly, $\beta_\varepsilon = \sigma_{F_\varepsilon} 1_n$ and $\beta_{q'} = \sigma_{F_{q'}} 1_n$, and the idiosyncratic shocks w_v, w_ε, and $w_{q'}$ are i.i.d. across assets with variances $\sigma^2_{w_v}$, $\sigma^2_{w_\varepsilon}$, and $\sigma^2_{w_{q'}}$, respectively, for each asset. In this case, the "market portfolio" $\beta := 1_n$ is the most inefficient portfolio for n sufficiently large (regardless of whether Assumption $1'$ is satisfied). As n grows without limit, the proportion η^β/η of total inefficiency accounted for by this portfolio approaches one.

It may be instructive to compare the inefficiency of the overall market with that of a

[24] To see this, note that (a) the variance of each (unit-norm) micro portfolio decreases as $1/n^2$ with n; (b) 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 (c) there are $n - k_v$-independent micro portfolios.

nondiversified portfolio. Specifically, we consider a portfolio consisting only of asset 1, which we denote by e_1. With our symmetry assumptions, the conditional variance matrices $\Sigma_{v|s}$ and $\Sigma_{v|p}$ can be computed explicitly, yielding the formulae for the inefficiency of asset 1 and the market, respectively:

$$\eta^{e_1} = \frac{1}{2}\log\left(\frac{1}{\tau^p(n)} + \frac{n-1}{\tau^p(0)}\right) - \frac{1}{2}\log\left(\frac{1}{\tau^s(n)} + \frac{n-1}{\tau^s(0)}\right), \tag{25}$$

$$\eta^{\beta} = \frac{1}{2}\log(\tau^s(n)) - \frac{1}{2}\log(\tau^p(n)). \tag{26}$$

The terms $\tau^p(\lambda)$ and $\tau^s(\lambda)$ are given as follows for $\lambda \in \{0, n\}$:[25]

$$\begin{aligned}\tau^p(\lambda) = &\left(\sigma^2_{w_v} + \sigma^2_{F_v}\lambda\right)^{-1} + \left(\sigma^2_{w_\varepsilon} + \sigma^2_{F_\varepsilon}\lambda\right)^{-1} \\ &\times \left(1 + \gamma^2 I^{-2} n^{-2}\left(\sigma^2_{w_\varepsilon} + \sigma^2_{F_\varepsilon}\lambda\right)\left(\sigma^2_{w_{q'}} + \sigma^2_{F_q}\lambda\right)\right)^{-1},\end{aligned} \tag{27}$$

$$\tau^s(\lambda) = \left(\sigma^2_{w_v} + \sigma^2_{F_v}\lambda\right)^{-1} + \left(\sigma^2_{w_\varepsilon} + \sigma^2_{F_\varepsilon}\lambda\right)^{-1}. \tag{28}$$

We have $\tau^p(0) \approx \tau^s(0)$ for large n, meaning that idiosyncratic information (corresponding to micro-portfolios, $\lambda = 0$) is reflected almost equally by the prices and the signals. This idiosyncratic information affects the efficiency of a single asset, but is diversified away for the index, that is, $\tau^p(0)$ and $\tau^s(0)$ feature in (25), but not in (26). Further, we see that $\eta^{e_1} < \eta^{\beta}$ from Equations (25) and (26) and the concavity of the logarithm. In words, asset 1 is less inefficient than the index. The economic intuition is that, while the single asset is more volatile than the index, the additional volatility is due to the idiosyncratic component, and the information about this component is (almost) entirely impounded in the price. Consequently, a larger proportion of the information available about the payoff of asset 1 is inferred from the price. This is precisely our definition of efficiency.

We illustrate these statements numerically. To that end, we consider a parametric model (described further in Appendix A) and choose the following parameters. There are $\bar{S} =$

[25]Here, $\tau^p(n)^{-1}$ and $\tau^s(n)^{-1}$ are the eigenvalues of $\Sigma_{v|p}$ and $\Sigma_{v|s}$, respectively, corresponding to the eigenvector β, and $\tau^p(0)^{-1}$ and $\tau^s(0)^{-1}$ are the eigenvalues corresponding to all other eigenvectors, that is, any vector orthogonal to β. We note that, under Assumption 1, matrices $\Sigma_{v|p}$, $\Sigma_{v|s}$, and G have the same eigenvectors.

10^8 investors, each with a relative risk aversion of $\gamma^R = 3$ and wealth of $W = \$100,000$. There are $\bar{M} = 4,000$ asset managers. The fundamentals have a one-factor structure, as described above. The factor volatility is $\sigma_{F_v} = 0.18W\bar{S}$ and the idiosyncratic volatility is $\sigma_{w_v} = 0.37W\bar{S}$. The noise has the same factor structure but with double the volatility of factors and idiosyncratic shocks. The number of securities is $n = 100$, with average supply $\bar{q}_i' = 1$ and supply uncertainty characterized by $\sigma^2_{F_{q'}} = 0.03$ and $\sigma^2_{w_{q'}} = 0.06$ for all $i \leq n$. The marginal cost of asset management is ${k_p}^{\%} = {k_a}^{\%} = 0.10\%$ and the cost of information is $k = 2e7$. The search cost is $c = 84\,(I/M)^{0.25}$. The cost of self-directed investment is zero for 10% of investors and uniformly distributed on $[0, 2\%]$ for the rest.

With these parameters, the equilibrium overall market inefficiency is $\eta = 6\%$, so the parameters are chosen to match the level of inefficiency of the calibration in Section 1.4. The inefficiency of the market portfolio, as given by Equation (26), is $\eta^\beta = 5.22\%$, which represents, indeed, a large part of the overall inefficiency. The inefficiency of a single stock, say stock 1, is computed using Equation (25) to be $\eta^{e_1} = 1.09\%$. This inefficiency is smaller than that of the market portfolio because an individual stock has both common and idiosyncratic components, as discussed before. The idiosyncratic component has variance more than four times larger[26] than the common one, but this component is virtually entirely revealed by prices. If an investor was restricted to trading in asset 1, then the use of information would only benefit her by 1.09% in terms of wealth certainty equivalent, since she would have to take on sizable idiosyncratic risk without any information advantage. Trading the market portfolio, in contrast, would confer a 5.22% benefit.

We also emphasize that a single stock is not a true micro portfolio in the sense used in our model since micro portfolios must have zero factor exposure. A true micro portfolio ζ is market neutral, $\zeta^\top \beta = 0$, and we denote the inefficiency of such a portfolio as the micro-inefficiency $\eta^\perp := \eta^\zeta$. In this example, $\eta^\perp = \frac{1}{2}\log(\tau^s(0)) - \frac{1}{2}\log(\tau^p(0)) = 0.0078\%$, so we see that the true micro-inefficiency is much lower than the inefficiency of an individual stock. This low micro-inefficiency reflects that relative prices are more "correct" than any

[26] We have $\frac{\sigma^2_{w_v}}{\sigma^2_{F_v}} = \left(\frac{0.37}{0.18}\right)^2 = 4.22$.

individual price.

Further, we can see how all these inefficiencies add up to the overall inefficiency. There are $n-1 = 99$ micro portfolios and one macro-inefficiency, and these add to the total inefficiency, $(n-1)\eta^{\perp} + \eta^{\beta} = 99 \times 0.0078\% + 5.22\% = 6\% = \eta$.

We can, of course, compute all of the equilibrium quantities in this example. For example, the proportion of informed investors is $I/\bar{S} = 0.194$ and there are $M = 1,152$ active managers.

Example 2: Two-factor model. Next, we consider a multifactor example. To provide some numbers and a graphical illustration, we modify the parameters in example 1 by adding a second factor, so that $\beta_v = \beta\sigma_{F_v}$ where $\beta = (\beta_1, \beta_2) \in \mathbb{R}^{n\times 2}$ now consists of two columns. Similarly, $\beta_\varepsilon = \beta\sigma_{F_\varepsilon}$ and $\beta_{q'} = \beta\sigma_{F_{q'}}$.

Here, factor 1 is the market factor (as before), $\beta_1 = 1_n$, which means that all assets are part of the market. Factor 2 is a long-short factor, $\beta_2 = (b, -b, b, -b, ..., (-1)^{n-1}b)^{\top}$. For example, we can think of factor 2 as the Fama-French value-minus-growth factor, HML. The alternating signs in factor 2 means that every other asset is a value stock, and the remaining ones are growth stocks. We select $b = 0.61$ to match the finding of Roll (1988) that "the mean R^2s were, respectively, .179 for the CAPM and .244 for the APT."

In this example, the eigenvalues of $\beta_v^{\top}\beta_v/n$ are $\sigma^2_{F_v}$ and $b^2\sigma^2_{F_v}$ for all n even, so we obtain that factor 2 is a weaker driver of returns since $|b| < 1$. For n large enough (and even), $\{\beta_1, \beta_2\}$ achieves maximal inefficiency among all pairs of portfolios. (Naturally, any other pair generating β_1 and β_2 also does.)

Figure 3 displays the proportion of the total inefficiency that is due to each of the two factors, as well as all of the micro portfolios, for a number of stocks ranging between 4 and 1000.

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, 80.8% of the overall inefficiency is due to the inefficiency of the expected market portfolio $\beta_1 = 1_n$, 18.0% is due to

the other systematic factor, namely, the relative-value portfolio $\beta_2 = (.61, -.61, ..., -.61)^\top$, and the remaining 1.2% 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 our example contains as many as 1,000 securities. In other words, micro-inefficiencies are diminished when informed investing virtually eliminates near-arbitrage opportunities, at least in the model. Hence, most of the inefficiency arises from the nondiversifiable risk due to the two factors. Most of the inefficiency is in the expected market portfolio, though a nontrivial 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, nontrivial mispricing can exist when many inefficient trades are correlated, consistent with the evidence that most return drivers indeed are based on factor structures of such correlated trades (see Kelly et al. 2018 and the references therein). In contrast, truly idiosyncratic mispricing should be minimal according to the model. This is a quantitative prediction that may or not stand the test of data, especially before transaction costs.[27]

Implications of investment management. Proposition 5 means that, with many assets, investors have 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 premiums (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

[27]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 does Roll (1988), whose study we use for our choice of parameters, so a real test of the model should use parameters consistent with the test portfolios.

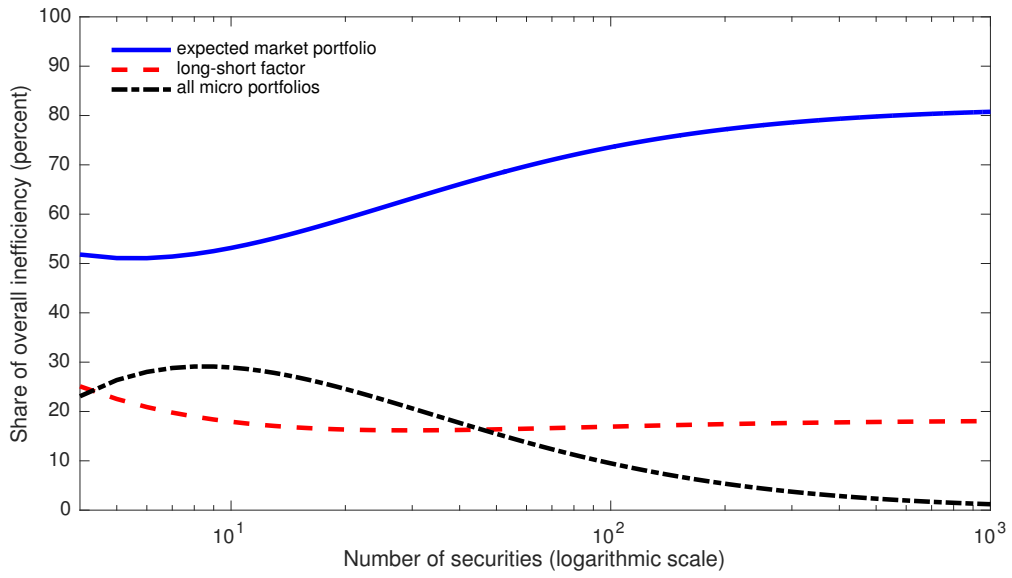

Figure 3: **Decomposing overall inefficiency**
This figure shows the share of overall market inefficiency arising from the expected market portfolio (solid line), long-short portfolios, such as the Fama and French (1993) factors, called high-minus-low (HML) and small-minus-big (SMB) used in much of the 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.

them. These predictions appear consistent with some empirical observations.[28]

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 (2018) estimates that "the factor industry is $1.9 trillion in AUM."

In summary, we derive endogenously the meaning of macro-efficiency (by looking for the

[28]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.

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. Our results complement those of Glasserman and Mamaysky (2018), who also provide conditions for higher macro-inefficiency with a single factor and an exogenous definition of macro versus micro information, but endogenous information choices. Finally, we note that our results on Samuelson's dictum would also apply in a model without asset managers in which investors choose individually whether or not to be informed. However, asset managers play a central role in real-world information acquisition and, further, asset managers are important for the next section, where we will discuss the implications of changing asset management costs.

3 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 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 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 have the implications conjectured by Samuelson's. In particular, when Samuelson's dictum holds for

the *levels* of inefficiency, then we can look at the relative *changes* in macro-inefficiency versus micro-inefficiency. Samuelson's dictum holds for the level of inefficiency under Assumption 1′ as shown in Proposition 4(b), or when n is not too small under Assumption 3 as stated in Proposition 5. Here, we focus on the latter case and impose Assumption 3 in the rest of this section. Recall that macro-inefficiency is given by η^{β_v} as in Proposition 5, while *total* micro-inefficiency is the residual to η.

Proposition 6 (Information costs and efficiency) *When the cost of information k decreases, overall asset price inefficiency η decreases and the macro-inefficiency η^{β_v} decreases by more than the total micro-inefficiency ($\eta - \eta^{\beta_v}$) as long as n is not too small. Further, the number of self-directed investors remains unchanged, the numbers of active investors I and of informed active managers M increase, 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, they remain 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 overall asset price inefficiency (η) increases and the macro-inefficiency η^{β_v} increases by more than the total micro-inefficiency ($\eta - \eta^{\beta_v}$) as long as n is not too small. Further, the number of passive investors S_p increases, the numbers of active investors I, self-directed investors, 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. As 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) find 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 further shows that such a rise in market inefficiency is greater for macro-inefficiency than micro-inefficiency, as macro-inefficiency tends to be more variable as discussed above. The proposition also makes predictions for fees, which we can compare with the 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. Propo-

sition 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 such a reduction need not obtain. To understand this feature, note the two effects at play: First, a lower cost of passive investing directly lowers the cost of active through competitive effects as seen in Equation (10). Second, having fewer active investors leads to a higher market inefficiency (η), which increases the value of active management, and hence the fee (the second term in Equation (10)). This second effect mitigates the reduction in the active fee (and can in some cases even reverse it).

Example 1, continued. Next, we illustrate some of the effects of changing costs of passive investing with an example. We use the parameters from example 1 in Section 2, but vary the cost of passive investing to illustrate the results predicted by Proposition 7.

Figure 4, panel A, shows how market inefficiency and active fees change when the cost of passive investing ${f_p}^{\%}$ varies. Figure 4, panel B, shows how the ownership structure depends on the passive fees in the numerical example. Naturally, a lower passive fee leads to less active investing as seen by moving from right to left in Figure 4, panel B. When there are fewer active investors, the market becomes less efficient, as seen by moving from right to left in Figure 4, panel A, and most of that change is due to the macro-inefficiency. (Remember also that the total micro-inefficiency equals $(n-1)\eta^{\perp} = 99\eta^{\perp}$, highlighting the small scale of the inefficiency of any one micro portfolio, $\eta^{\perp}$, and of its sensitivity to parameters.)

A change in the cost of passive investing actually does not change the active management fee in the numerical example, as seen in Figure 4, panel A. This result happens because, as discussed above, a reduction in passive fees both (a) lowers the active fee via competitive pressure and (b) increases the active fee via higher market inefficiency. These two effects exactly offset each other in this example. (The fact that the offset is exact, however, is not a general property of the model.)

Figure 4, panel B, also shows how lower passive costs imply more passive asset management, less active investment, and less self-directed investment. These findings can be

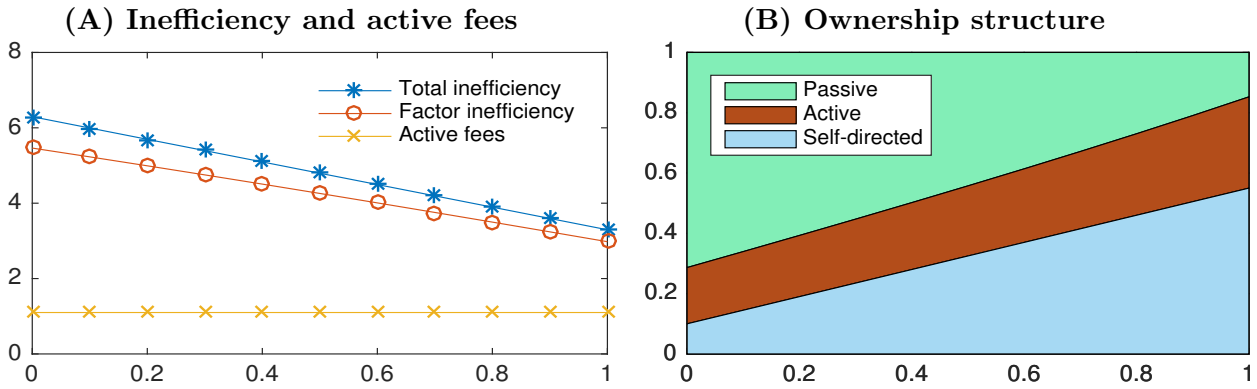

Figure 4: **Changing passive investment costs**
This figure shows properties of the model implied by different values of the percentage cost of passive investment, $f_p^{\%}$ (on the x-axis, in percent). Panel A shows the total inefficiency (η), the macro-inefficiency (η^{β}), and the active fee ($f_a^{\%}$), all in percent. Panel B shows the fractions of ownership allocated to passive management (S_p, upper area in the figure), active management (I, middle area), and self-directed investment (bottom area).

viewed as the model-based counterpart to some of the recent trends in real-world markets seen in Figure 1. Since the mid-1990s, passive management has grown as passive fees have declined, self-directed investment has fallen (although this trend changes in the end of the sample, perhaps because of reductions in the cost of self-directed investment from discount brokerages), and, in the last part of this time period, active management has also starting to reverse its prior multidecade growth.

4 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 natural 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 optimal 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 indexes should emerge as the most successful.

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.

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.

Bai, J., T. Philippon, and A. Savov (2016). Have financial markets become more informative? *Journal of Financial Economics 122*(3), 625–654.

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.

Biais, B., P. Bossaerts, and C. Spatt (2010). Equilibrium asset pricing and portfolio choice under asymmetric information. *The Review of Financial Studies 23*(4), 1503–1543.

BlackRock (2018). Factor investing: 2018 landscape. Technical report.

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 (2020). Asset management contracts and equilibrium prices. *London School of Economics, 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.

Dávila, E. and C. Parlatore (2018). Identifying price informativeness. Technical report, National Bureau of Economic Research.

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.

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.

Peress, J. (2005). Information vs. entry costs: What explains u.s. stock market evolution? *Journal of Financial and Quantitative Analysis 40*(3), 563–594.

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.

Xiao, Y., H. Yan, and J. Zhang (2021). A global version of samuelson's dictum. *Available at SSRN 3810241*.

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:

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

where $\alpha>0$ and $\bar{c}>0$ are parameters. Combining this with (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

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

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

Given (7), (8), (10), and (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, so we take I as given. The optimal uninformed demand x_u and informed demand x_i can be written using the notation above as

$$x_u=\frac{1}{\gamma}\Sigma_{v|p}^{-1}\left(\text{E}[v|p]-p\right) \tag{A.5}$$

$$x_i=\frac{1}{\gamma}\Sigma_{v|s}^{-1}\left(\text{E}[v|s]-p\right). \tag{A.6}$$

To compute the relevant expectations and variance, we conjecture the form (3) 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:

$$\mathrm{E}[v|p] = \mathrm{E}[v|\hat{p}] = \bar{v} + \beta_{v,\hat{p}}\hat{p} = \bar{v} + \Sigma_v \Sigma_{\hat{p}}^{-1}\hat{p} \tag{A.7}$$

$$\mathrm{E}[v|s] = \mathrm{E}[v|v+\varepsilon] = \bar{v} + \beta_{v,s}(s-\bar{v}) = \bar{v} + \Sigma_v \Sigma_s^{-1}(s-\bar{v}) \tag{A.8}$$

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

$$\Sigma_{v|s} = \left(\Sigma_v^{-1} + \Sigma_\varepsilon^{-1}\right)^{-1} = \Sigma_v \left(\Sigma_v + \Sigma_\varepsilon\right)^{-1} \Sigma_\varepsilon. \tag{A.10}$$

We can now insert the demands (A.5) and (A.6) into the market-clearing condition (2), which yields 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 for these variables must equal zero, and, similarly, the constant term must be zero, yielding three equations. Solving these three equations leads to the equilibrium coefficients in the price function (3):

$$\theta_0 = \bar{v} - \gamma \left(U \left(\Sigma_{v|p}\right)^{-1} + I \left(\Sigma_{v|s}\right)^{-1}\right)^{-1} \bar{q} \tag{A.11}$$

$$\theta_q = \frac{\gamma}{I}\Sigma_\varepsilon \tag{A.12}$$

$$\theta_s = \left(U \left(\gamma\Sigma_{v|p}\right)^{-1} + I \left(\gamma\Sigma_{v|s}\right)^{-1}\right)^{-1} \left(\theta_q^{-1} + U \left(\gamma\Sigma_{v|p}\right)^{-1} \Sigma_v \Sigma_{\hat{p}}^{-1}\right). \tag{A.13}$$

Inserting the expression for θ_q from (A.12) into (A.9) and using that $\Sigma_{\varepsilon+\theta_q q} = \Sigma_\varepsilon + \theta_q \Sigma_q \theta_q^\top$ yields

$$\Sigma_{v|p} = \left(\Sigma_v^{-1} + (\Sigma_\varepsilon + \gamma^2 I^{-2} \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1}\right)^{-1}. \tag{A.14}$$

Finally, inserting (A.14) and (A.10) into the definition of market inefficiency (6) shows how market inefficiency depends on the number of active investors I, that is, Equation (9). The part of the equilibrium pertaining to investors' choice of the modality of investment, the managers' information choice, and fee setting is derived in the body of the paper. We present here one argument not given in the text, supporting the claim that no investor acquires information to invest on her own. If she did, her utility would be $W + u_i - k - d_l$, while with an active manager she would achieve $W + u_i - c(M,I) - f_a$. The net benefit of choosing a manager is

$$\begin{aligned} k - c(M,I) - f_a + d_l &= k - c(M,I) - \frac{k_a + k_p}{2} - \frac{\eta}{2\gamma} + d_l \\ &= k - \frac{\eta}{2\gamma} + \frac{k_a - k_p}{2} - \frac{k_a + k_p}{2} - \frac{\eta}{2\gamma} + d_l \\ &= k - k_p - 2k\frac{M}{I} + (k_p - k_a) + d_l \\ &= k\left(1 - 2\frac{M}{I}\right) - k_a + d_l, \end{aligned} \tag{A.15}$$

where we used (10), (12), and (11) in succession. We argue that, for reasonable parameters, (A.15) is positive, implying that the investor prefers an active manager to collecting information by oneself. This conclusion arises for two reasons. First, each fund services (many) more than two investors, so the ratio $2M/I$ is (far) below one, implying that the first term is positive. Second, the manager's marginal cost is likely less than the investor's own marginal cost, so $-k_a + d_l \geq 0$ and, even if this is not the case, then these marginal costs are swamped by the fixed infrastructure cost of information ($k \gg k_a$), such that (A.15) is positive overall.

A.4 Further Results

A.4.1 Active portfolio holdings

The following proposition characterizes the portfolio holdings of the informed investors.

Proposition 8 (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." 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)).

A.4.2 Change in noise

Section 3 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 noise trading is primarily due 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* I *and* M *and 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 informed active managers. The effect on market inefficiency is ambiguous; it depends on how many patsies (noise traders) versus sharks (informed investors) have left the "poker table."

A.4.3 Efficiency, entropy, and the value of information

We can also 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 so 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 has been studied in a number of papers in the literature, and the following proposition further establishes a link to the degree of market inefficiency.[29]

Proposition 1 (Extended version) *Consider a set of portfolios collected in the matrix* ζ*. The following quantities are equal to each other.*

(a) *the inefficiency of the portfolio collection* ζ*,* η^{ζ};

(b) *the utility gain to an investor becoming informed when restricted to* ζ*,* $(u_i - u_u)\gamma$;

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

[29]See also Admati and Pfleiderer (1987), who links the value of information to a ratio of determinants that coincides with our definition of inefficiency, Veldkamp (2011) for an overview of research on information choice, and Cabrales et al. (2013) for a recent contribution on entropy as the economic value of information and for further references.

(d) *the expected Kullback-Leibler divergence, KL, of the distribution of* $\zeta^\top v$ *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 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.

A.5 Proofs

Some of our results follow from a more general result, Lemma A.5, which we state and prove below. For clarity, we spell out formally the assumption on which this lemma relies.

Assumption 1″ *Fundamentals have a factor structure:*

$$v = \bar{v} + \beta F_v + w_v \tag{A.16}$$
$$\varepsilon = \beta F_\varepsilon + w_\varepsilon \tag{A.17}$$
$$q = \bar{q} + \beta F_q + w_q, \tag{A.18}$$

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* $\sigma^2_{w_v}$, $\sigma^2_{w_\varepsilon}$, *and* $\sigma^2_{w_q}$, *respectively, for each asset.*

We use the following notation:

$$G = \Sigma_{v|s}^{-\frac{1}{2}} \Sigma_{v|p} \Sigma_{v|s}^{-\frac{1}{2}} \tag{A.19}$$
$$O = \beta\beta^\top. \tag{A.20}$$

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

$$g(\lambda) \equiv 1 + \frac{X(\lambda)Y(\lambda)}{1 + X(\lambda) + Y(\lambda)}, \tag{A.21}$$

with

$$X(\lambda) \equiv \frac{\sigma_{w_v}^2 + \sigma_{F_v}^2 \lambda}{\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2 \lambda} \tag{A.22}$$

$$Y(\lambda) \equiv \gamma^2 I^{-2} \left(\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2 \lambda\right) \left(\sigma_{w_q}^2 + \sigma_{F_q}^2 \lambda\right). \tag{A.23}$$

If Assumption 1′ holds, then g is an increasing function. **Proof of Lemma A.5.** We start by noting that the variance matrices of the fundamental quantities v, ε, and q have the form $a\mathrm{I}_n + bO$ for appropriate (positive) scalars a and b. E.g.,

$$\Sigma_v = \sigma_{w_v}^2 \mathrm{I}_n + \sigma_{F_v}^2 O. \tag{A.24}$$

We need to work with a slightly more general set of matrices. Specifically, with $B = \beta^\top \beta$, we consider matrices of the form $a\mathrm{I}_n + \beta f(\beta^\top \beta)\beta^\top$, where $f : \mathbb{R}_+ \to \mathbb{R}$ is a finite-valued function.[30] 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

$$\left(\mathrm{I}_n + \beta f(\beta^\top \beta)\beta^\top\right)^{-1} = \mathrm{I}_n + \beta \hat{f}(\beta^\top \beta)\beta^\top \tag{A.25}$$

$$= \mathrm{I}_n - \beta(\mathrm{I}_n + \beta^\top \beta f(\beta^\top \beta))^{-1} f(\beta^\top \beta)\beta^\top, \tag{A.26}$$

with

$$\hat{f}(x) = -\frac{f(x)}{1 + x f(x)} \tag{A.27}$$

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 $a\mathrm{I}_n + \beta f(B)\beta^\top$. (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 $a\mathrm{I}_n + \beta f(B)\beta^\top$ 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 $M_1 = a_1 \mathrm{I}_n + \beta f_1(B)\beta^\top$ and $M_2 = a_2 \mathrm{I}_n + \beta f_2(B)\beta^\top$ and the bivariate function $F(x, y)$ being 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

$$G = \Sigma_{v|s}^{-\frac{1}{2}} \Sigma_{v|p} \Sigma_{v|s}^{-\frac{1}{2}} = \Sigma_{v|s}^{-1} \Sigma_{v|p} = \left(\Sigma_v^{-1} + \Sigma_\varepsilon^{-1}\right) \left(\Sigma_v^{-1} + (\Sigma_\varepsilon + \theta_q \Sigma_q \theta_q)^{-1}\right)^{-1}. \tag{A.28}$$

[30] 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 $Uf(D)U^{-1}$.

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.28). Specifically, we have

$$g(\lambda) = \frac{\left(\sigma_{w_v}^2 + \sigma_{F_v}^2\lambda\right)^{-1} + \left(\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2\lambda\right)^{-1}}{\left(\sigma_{w_v}^2 + \sigma_{F_v}^2\lambda\right)^{-1} + \left(\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2\lambda + \gamma^2 I^{-2}\left(\sigma_{w_\varepsilon}^2 + \sigma_{F_\varepsilon}^2\lambda\right)^2\left(\sigma_{w_q}^2 + \sigma_{F_q}^2\lambda\right)\right)^{-1}}. \tag{A.29}$$

A simple manipulation allows one to check that Equation (A.21) 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. Y is clearly increasing, while X increases if and only if Assumption 1′ holds. ■

Proof of Proposition 1 (Extended version). Let us start by defining $\tilde{v} = \zeta^\top v$ and $\tilde{p} = \zeta^\top p$. Clearly, if $\zeta = \mathrm{I}_n$, then we obtain the statements for the entire market, covering the first part of Proposition 1. We need to calculate the utilities u_u and u_i, and we use the formula

$$\mathrm{E}\left[e^{x^\top Ax + b^\top x}\right] = \det(\mathrm{I} - 2\Omega A)^{-\frac{1}{2}} e^{\frac{1}{2}b^\top(\mathrm{I}-2\Omega A)^{-1}\Omega b} \tag{A.30}$$

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

$$\mathrm{E}\left[e^{-\gamma(x_i(\tilde{v}-\tilde{p}))}\,|p\right] = \mathrm{E}\left[e^{-\frac{1}{2}(\mathrm{E}[\tilde{v}|s]-\tilde{p})^\top\Sigma_{\tilde{v}|s}^{-1}(\mathrm{E}[\tilde{v}|s]-\tilde{p})}|p\right] \tag{A.31}$$

by letting $x = \mathrm{E}[\tilde{v}|s] - \mathrm{E}[\tilde{v}|p]$, $A = -\frac{1}{2}\Sigma_{\tilde{v}|s}^{-1}$, and $b^\top = (\mathrm{E}[\tilde{v}|p] - \tilde{p})^\top \Sigma_{\tilde{v}|s}^{-1}$ to evaluate (A.31) as

$$\begin{aligned}
&\mathrm{E}\left[e^{x^\top Ax + b^\top x - \frac{1}{2}(\mathrm{E}[\tilde{v}|p]-\tilde{p})^\top\Sigma_{\tilde{v}|s}^{-1}(\mathrm{E}[\tilde{v}|p]-\tilde{p})}|p\right] \\
&= \det(\mathrm{I}_n + \Omega\Sigma_{\tilde{v}|s}^{-1})^{-\frac{1}{2}} \\
&\quad \times e^{\frac{1}{2}(\mathrm{E}[\tilde{v}|p]-\tilde{p})^\top\Sigma_{\tilde{v}|s}^{-1}(\mathrm{I}+\Omega\Sigma_{\tilde{v}|s}^{-1})^{-1}\Omega\Sigma_{\tilde{v}|s}^{-1}(\mathrm{E}[\tilde{v}|p]-\tilde{p}) - \frac{1}{2}(\mathrm{E}[\tilde{v}|p]-\tilde{p})^\top\Sigma_{\tilde{v}|s}^{-1}(\mathrm{E}[\tilde{v}|p]-\tilde{p})}
\end{aligned} \tag{A.32}$$

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

$$\gamma(u_i - u_u) = \frac{1}{2}\left(\log(\det(\Sigma_{\tilde{v}|p})) - \log(\det(\Sigma_{\tilde{v}|s}))\right), \tag{A.33}$$

which shows the equivalence of utility gain with inefficiency. In turn, the latter is the same as the entropy difference for the normal distributions of $\tilde{v}$ conditional on p and s, respectively. Thus, the quantities given in parts (a), (b), and (c) are equal. We go further by using the fact that the Kullback-Leibler divergence of a n-dimensional multivariate normal distribution

with mean μ_1 and variance Σ_1 from one with mean μ_0 and variance Σ_0 is

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

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

Proof of Proposition 3. (a) Start with the market clearing condition

$$\begin{aligned} q &= Ux_u + Ix_i \\ &= U\left(\gamma\Sigma_{v|p}\right)^{-1}\left(\mathrm{E}[v|p]-p\right) + I\left(\gamma\Sigma_{v|s}\right)^{-1}\left(\mathrm{E}[v|s,p]-p\right). \end{aligned} \tag{A.34}$$

Take expectations conditional on p and rewrite to get

$$\mathrm{E}[q|p] = \left(U\mathrm{I}_n + I\Sigma_{v|s}^{-1}\Sigma_{v|p}\right)x_u. \tag{A.35}$$

Solving for x_u yields Equation (18). Let us use, in this proof, the notation $A \sim B$ for two matrices that are scalar multiples of each other. To see the implications of the sufficient condition $\Sigma_v \sim \Sigma_\varepsilon \sim \Sigma_q^{-1}$, that is, Assumption 2, we note the following:

$$\theta_q \sim \Sigma_\varepsilon \tag{A.36}$$

$$\theta_s \sim \mathrm{I}_n \tag{A.37}$$

$$\Sigma_{v|s} = \Sigma_v\left(\Sigma_v+\Sigma_\varepsilon\right)^{-1}\Sigma_\varepsilon \sim \Sigma_\varepsilon \tag{A.38}$$

$$\Sigma_{v|p} = \Sigma_v\left(\Sigma_v+\Sigma_\varepsilon+\theta_q\Sigma_q\theta_q\right)^{-1}\left(\Sigma_\varepsilon+\theta_q\Sigma_q\theta_q\right) \sim \Sigma_\varepsilon \tag{A.39}$$

Consequently, $\Sigma_{v|s}^{-1}\Sigma_{v|p}$ is a scalar and x_u is proportional to $\mathrm{E}[q|p]$. 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 $(U\mathrm{I}_n + I\Sigma_{v|s}^{-1}\Sigma_{v|p})^{-1}$ is positive definite. Further, the assumptions of Lemma A.5 are satisfied. It follows that $H \equiv (U\mathrm{I}_n + I\Sigma_{v|s}^{-1}\Sigma_{v|p})^{-1}$ takes the form $a_0\mathrm{I}_n - a_1 O$. 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) The result under Assumption 2 or 1 follows from part 1 of the proposition by taking unconditional expectations. Under Assumption 1′, we use Equation (19), taking into account the sign of A_1 and the normalization of β to derive $E(x_{u,i}) \leq E(x_{u,j})$. Finally, under the independence assumptions stated in the last scenario, 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 4. (a) Under Assumption 2, G is a scalar (see proof of Proposition 3, part 1a), and thus all portfolios have the same inefficiency.

(b) We note that the market portfolio, β, is the only nonzero 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 A.5 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 5. Before we begin the proof proper, let us make the convention that scalars are to be construed as the appropriate multiples of the identity matrix. This is in keeping with our thinking of all matrices as operators between Euclidean spaces. In particular, we are going to consider the operator norm. We remind the reader that, for a matrix $A \in \mathbb{R}^{m\times n}$, this is defined as $\|A\| = \max_{x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|}$. For a vector in $\mathbb{R}^n$, the operator norm coincides with the Euclidean norm. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the norm is equal to the largest eigenvalue of A, while for a general matrix A, $\|A\|^2 = \|A^\top A\| = \|AA^\top\|$. We write $A > 0$ if A is positive definite; naturally, $A > B$ is equivalent to $A - B > 0$. Finally, as we wrote in the text, we suppress notational dependence on n, while all statements are meant as limit statements as n grows large. We hereafter use the notation $\sim$ instead of "big O." Both APT results involve sequences of economies, in which the functions $\eta_n(\cdot)$ are different and therefore so are the (largest) equilibrium values of I_n. As the proof establishes, the sequence $\eta_n(\cdot)$ converges pointwise to a continuous function. In addition, the convergence is uniform on compacts. The sequence I_n therefore has a limit I, which satisfies the limiting equations defining equilibria in finite economies. The same uniform convergence obtains for $\eta_n^{\beta_v}(\cdot)$ and the APT pricing error of any given asset. As a consequence, any limit statement that holds for all I in a compact set also holds when applied to the convergent sequence I_n. The proof therefore fixes I, but one can just as easily replace it with I_n. We also introduce the notation D_x, for $x \in \{v, \varepsilon, q\}$, to refer to the diagonal matrix of variances of the idiosyncratic components of x.

[APT of returns.] The unconditional risk premium is given by (A.11):

$$\begin{aligned}
\mathrm{E}[v-p] &= \gamma\left(U\left(\Sigma_{v|p}\right)^{-1} + I\left(\Sigma_{v|s}\right)^{-1}\right)^{-1}\bar{q} \\
&= \gamma\left(U\left(\Sigma_v\Sigma^{-1}_{v+\varepsilon+\theta_q q}\Sigma_{\varepsilon+\theta_q q}\right)^{-1} + I\left(\Sigma_v\Sigma^{-1}_{v+\varepsilon}\Sigma_\varepsilon\right)^{-1}\right)^{-1}\bar{q} \\
&= \gamma\Sigma_v\left(U\Sigma^{-1}_{\varepsilon+\theta_q q}\Sigma_{v+\varepsilon+\theta_q q} + I\Sigma^{-1}_{\varepsilon}\Sigma_{v+\varepsilon}\right)^{-1}\bar{q} \\
&= \gamma\left(\beta_v\beta_v^\top + D_v\right)\left(U\Sigma^{-1}_{\varepsilon+\theta_q q}\Sigma_{v+\varepsilon+\theta_q q} + I\Sigma^{-1}_{\varepsilon}\Sigma_{v+\varepsilon}\right)^{-1}\bar{q}. \qquad \text{(A.40)}
\end{aligned}$$

The above term in $\beta_v\beta_v^\top$ is in agreement with the return APT. We still need confirm that the risk premiums λ implied by this expression have a well defined limit, and then to show that the second term has (uniformly) bounded Euclidian norm. Letting $u = U/(U+I)$ and

$i = 1 - u$, we have

$$\begin{aligned}
&\left(U\Sigma^{-1}_{\varepsilon+\theta_q q}\Sigma_{v+\varepsilon+\theta_q q} + I\Sigma^{-1}_{\varepsilon}\Sigma_{v+\varepsilon}\right)^{-1} \\
&\quad = (U+I)^{-1}\left(u\Sigma^{-1}_{\varepsilon+\theta_q q}\Sigma_{v+\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon}\Sigma_{v+\varepsilon}\right)^{-1} \\
&\quad = (U+I)^{-1}\left((u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} + \Sigma_v\right)^{-1}(u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} \\
&\quad = (U+I)^{-1}\left(1 - \left((u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} + \Sigma_v\right)^{-1}\Sigma_v\right).
\end{aligned} \tag{A.41}$$

By assumption, $\beta_v^\top \bar{q}$ has a well-defined limit. It remains to consider the term

$$\beta_v^\top \left((u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} + \Sigma_v\right)^{-1}(\beta_v\beta_v^\top + D_v)\bar{q}. \tag{A.42}$$

First, we evaluate

$$\beta_v^\top \left((u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} + \Sigma_v\right)^{-1}\beta_v \tag{A.43}$$

by noting that it can be bounded both above and below by similar expressions in which all instances of the diagonal matrices D_v, D_ε, and D_q are replaced by scalars, and furthermore $\Sigma_{\theta_q q}$ can be replaced with $s + \bar{\gamma}^2\beta_\varepsilon\beta_\varepsilon^\top\beta_q\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top$ for a scalar s, where $\bar{\gamma} := \gamma/I$. We now show that the limit does not depend on the values of the scalars. Indeed, using a "push-through" lemma (see Lemma A.5) and $\beta_v = \beta_\varepsilon a$, we have

$$\begin{aligned}
&\beta_v^\top \left((u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} + \Sigma_v\right)^{-1}\beta_v \\
&\quad \to B_{v\varepsilon}\left((u(B_\varepsilon + \bar{\gamma}^2 B_\varepsilon B_{\varepsilon q'} B^\top_{\varepsilon q'})^{-1} + iB_\varepsilon^{-1})^{-1} + aa^\top B_\varepsilon\right)^{-1} a.
\end{aligned} \tag{A.44}$$

Second, we note that

$$\beta_v^\top \left((u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} + \Sigma_v\right)^{-1} D_v\bar{q} \tag{A.45}$$

tends to zero in norm because $D_v\bar{q}$ is bounded while, by Lemma A.5,

$$\beta_v^\top \left((u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} + \Sigma_v\right)^{-1} \sim n^{-\frac{1}{2}} \tag{A.46}$$

provided that the matrix

$$A \equiv (u\Sigma^{-1}_{\varepsilon+\theta_q q} + i\Sigma^{-1}_{\varepsilon})^{-1} \tag{A.47}$$

has k_ε eigenvalues that increase linearly with n, while the remaining $n - k_\varepsilon$ are uniformly

bounded. To verify this condition, we write

$$\begin{aligned} A &= (u(\Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} + i\Sigma_\varepsilon^{-1})^{-1} \\ &= \Sigma_\varepsilon^{\frac{1}{2}} (u(1 + \bar{\gamma}^2 \Sigma_\varepsilon^{\frac{1}{2}} \Sigma_q \Sigma_\varepsilon^{\frac{1}{2}})^{-1} + i)^{-1} \Sigma_\varepsilon^{\frac{1}{2}}. \end{aligned} \tag{A.48}$$

The matrix in the outer parentheses immediately above is bounded below away from zero and above, given that $\Sigma_\varepsilon^{\frac{1}{2}} \Sigma_q \Sigma_\varepsilon^{\frac{1}{2}} > 0$ is bounded above in norm. Consequently $0 < \underline{m}' < \overline{m}' \in \mathbb{R}$ exist such that $\underline{m}' \Sigma_\varepsilon < A < \overline{m}' \Sigma_\varepsilon$. The conclusion about the eigenvalues of A follows by Weyl's theorem for Hermitian matrices, which states that the k-th highest eigenvalue of the larger matrix is larger than k-th highest eigenvalue of the other matrix. We therefore conclude that the risk premiums have well-defined limits. It remains to show that the pricing errors are bounded (in the sense of ℓ^2 norm). To apply Lemma A.5, we order the eigenvalues of A in decreasing order, and write

$$\begin{aligned} A &= V_A D_A V_A^\top \\ &= V_A (d_{k_v} + (1 - d_{k_v}) D_A (1 - d_{k_v})) V_A^\top + V_A d_{k_v} (D_A - d_{k_v}) d_{k_v} V_A^\top. \end{aligned} \tag{A.49}$$

with D_A the diagonal matrix having the k-th eigenvalue in k-th place and V_A the (orthonormal) matrix whose columns are eigenvectors, and d_{k_v} the a diagonal matrix with 1 in places 1 through k_v and zeros elsewhere. Expression (A.49) is the decomposition desired. Since $\Sigma_v = D_v + \beta_v \beta_v^\top$, Lemma A.5 shows that the second term in (A.40), namely,

$$\gamma D_v \Big(U \Sigma_{\varepsilon + \theta_q q}^{-1} \Sigma_{v+\varepsilon+\theta_q q} + I \Sigma_\varepsilon^{-1} \Sigma_{v+\varepsilon} \Big)^{-1} \bar{q}, \tag{A.50}$$

is bounded. We note that this is the same conclusion as reached by the classical APT. Under a stronger condition that $\bar{q}$ be uniformly bounded in the Euclidean norm, namely, that its Euclidean norm tend to zero, the errors tend to zero in norm.

[APT of inefficiency.] We proceed in two steps. First, we show that the total inefficiency in the market of size n is the same, in the limit as n grows to infinity, as that of k_v fictitious assets that pay F_v. Then we show that portfolios given by the columns of β_v, and by extension by any k_v portfolios whose projections on these columns are linearly independent, have the same inefficiency in the limit. In a second step, we show that the maximum inefficiency of any portfolio ζ such that $\beta_v \zeta = 0$ goes to zero with n. In this sense, any such portfolio has zero inefficiency in the limit. We start by expressing the inefficiency as

$$\begin{aligned} e^{2\eta} &= \frac{\det(\Sigma_{v|p})}{\det(\Sigma_{v|s})} = \frac{\det\left(\Sigma_v (\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} (\Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)\right)}{\det\left(\Sigma_v (\Sigma_v + \Sigma_\varepsilon)^{-1} \Sigma_\varepsilon\right)} \\ &= \frac{\det\left(\Sigma_\varepsilon^{-1} (\Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)\right)}{\det\left((\Sigma_v + \Sigma_\varepsilon)^{-1} (\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)\right)}. \end{aligned} \tag{A.51}$$

We now turn to the informational inefficiency of prices in conveying signals about F_v. In

other words, with an abuse of notation, we want to estimate

$$
\begin{aligned}
e^{2\eta^{F_v}} &\equiv \frac{\det(\mathrm{var}(F_v|\bar{p}))}{\det(\mathrm{var}(F_v|s))} && \text{(A.52)}\\
&= \frac{\det(1-\beta_v^\top(\Sigma_v+\Sigma_\varepsilon+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1}\beta_v)}{\det(1-\beta_v^\top(\Sigma_v+\Sigma_\varepsilon)^{-1}\beta_v)} && \text{(A.53)}\\
&= \frac{\det((D_v+\Sigma_\varepsilon+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)(\Sigma_v+\Sigma_\varepsilon+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1})}{\det((D_v+\Sigma_\varepsilon)(\Sigma_v+\Sigma_\varepsilon)^{-1})} && \text{(A.54)}\\
&= \frac{\det((D_v+\Sigma_\varepsilon)^{-1}(D_v+\Sigma_\varepsilon+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon))}{\det((\Sigma_\varepsilon+\Sigma_v)^{-1}(\Sigma_\varepsilon+\Sigma_v+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon))}. && \text{(A.55)}
\end{aligned}
$$

To infer $\eta^{F_v}-\eta\to 0$, we only need to check that

$$
\det((D_v+\Sigma_\varepsilon)^{-1}(D_v+\Sigma_\varepsilon+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon))-\det(1+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q))\to 0. \tag{A.56}
$$

(The denominator above is clearly larger than one and thus is bounded below.) The first term equals

$$
\begin{aligned}
&\det(1+\bar{\gamma}^2(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)\\
&\quad=\det(1+\bar{\gamma}^2\Sigma_\varepsilon(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\Sigma_q) && \text{(A.57)}\\
&\quad\geq\det(1+\bar{\gamma}^2\beta_q^\top\Sigma_\varepsilon(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\beta_q).
\end{aligned}
$$

We note that

$$
\begin{aligned}
&\bar{\gamma}^2\beta_q^\top\Sigma_\varepsilon\beta_q-\bar{\gamma}^2\beta_q^\top\Sigma_\varepsilon(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\beta_q\\
&\qquad=\bar{\gamma}^2\beta_q^\top\Sigma_\varepsilon(\Sigma_\varepsilon^{-1}-(D_v+\Sigma_\varepsilon)^{-1})\Sigma_\varepsilon\beta_q\\
&\qquad=\bar{\gamma}^2\beta_q^\top\Sigma_\varepsilon(D_v+\Sigma_\varepsilon)^{-1}D_v\beta_q && \text{(A.58)}
\end{aligned}
$$

with $\|D_v\|\sim n^0$, $\|\beta_q\|\sim n^{-\frac{1}{2}}$, $\|\Sigma_\varepsilon^{\frac{1}{2}}\|\sim n^{\frac{1}{2}}$, and $\|(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon^{\frac{1}{2}}\|\sim n^0$,[31] which can be seen by considering

$$
\begin{aligned}
\|(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon^{\frac{1}{2}}\|^2&=\|(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon(D_v+\Sigma_\varepsilon)^{-1}\|\\
&\leq\|(D_v+\Sigma_\varepsilon)^{-1}(D_v+\Sigma_\varepsilon)(D_v+\Sigma_\varepsilon)^{-1}\|\\
&\leq(\min_i D_{vii})^{-1}. && \text{(A.59)}
\end{aligned}
$$

Thus, the right-hand side in (A.58) is of order $n^{-\frac{1}{2}}$. It consequently follows that, in the limit, (A.57) is bounded below by

$$
\det(1+\bar{\gamma}^2\Sigma_\varepsilon(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\Sigma_q)\geq\det(1+\bar{\gamma}^2\beta_q^\top\Sigma_\varepsilon\beta_q), \tag{A.60}
$$

[31] In fact, by Lemma A.5 the stronger statement $\|(D_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\|\sim n^0$ holds, rendering the product of order n^{-1}.

given that the matrix products in (A.58) have fixed dimension (k_q). For an upper bound, let $d_q = \|D_q\|$ and write

$$\begin{aligned}
&\det(1 + \bar{\gamma}^2(D_v + \Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon) \\
&\quad \leq \det(1 + \bar{\gamma}^2(D_v + \Sigma_\varepsilon)^{-1}\Sigma_\varepsilon(d_q + \beta_q\beta_q^\top)\Sigma_\varepsilon) \\
&\quad \leq \det(1 + \bar{\gamma}^2(d_q + \beta_q\beta_q^\top)\Sigma_\varepsilon). \\
&\quad \leq \det(1 + \bar{\gamma}^2(d_q + \beta_q\beta_q^\top)(d_\varepsilon + \beta_\varepsilon\beta_\varepsilon^\top)).
\end{aligned} \tag{A.61}$$

We will argue immediately below that

$$\det(1 + \bar{\gamma}^2(d_q + \beta_q\beta_q^\top)(d_\varepsilon + \beta_\varepsilon\beta_\varepsilon^\top)) \to \det(1 + \bar{\gamma}^2\beta_\varepsilon^\top\beta_q\beta_q^\top\beta_\varepsilon), \tag{A.62}$$

where the limit on the right-hand side is well defined by assumption, and it is equal to the limit of the numerator in (A.51). The matrix on the left-hand side of (A.62) is an $n \times n$ matrix for which $1 + \bar{\gamma}^2 d_q d_\varepsilon$ is an eigenvalue of multiplicity at least $n - k_v - k_\varepsilon$: any vector orthogonal to both β_ε and β_q is an eigenvector with this eigenvalue. It is important to note that $\bar{\gamma}^2 d_q d_\varepsilon \sim n^{-2}$, which implies

$$(1 + \bar{\gamma}^2 d_q d_\varepsilon)^n \to 1. \tag{A.63}$$

As for the nonzero eigenvalues of the low-rank matrix $d_q\beta_\varepsilon\beta_\varepsilon^\top + d_\varepsilon\beta_q\beta_q^\top + \beta_q\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top$, we note first that the norms of the first two terms are of order n^{-1}, while the third is of order n^0. In the limit, therefore, the determinant of interest, which by is given by (A.63) equals

$$\det(1 + \bar{\gamma}^2(d_q\beta_\varepsilon\beta_\varepsilon^\top + d_\varepsilon\beta_q\beta_q^\top + \beta_q\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top)) \tag{A.64}$$

tends to the limit of

$$\det(1 + \bar{\gamma}^2\beta_q\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top). \tag{A.65}$$

We have finally established that the lower bound in (A.60) is also an upper bound for the limit of the numerator of (A.55), which therefore equals that of the numerator of (A.51). Since (A.51) and (A.55) have the same denominator, η and η_{F_v} are equal (in the limit). Finally, we also estimate

$$\begin{aligned}
\eta^{\beta_v} &= \frac{\det(\beta_v^\top \mathrm{var}(v|\bar{p})\beta_v)}{\det(\beta_v^\top \mathrm{var}(v|\bar{s})\beta_v)} \\
&= \frac{\det(\beta_v^\top\Sigma_v\beta_v - \beta_v^\top\Sigma_v(\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1}\Sigma_v\beta_v)}{\det(\beta_v^\top\Sigma_v\beta_v - \beta_v^\top\Sigma_v(\Sigma_v + \Sigma_\varepsilon)^{-1}\Sigma_v\beta_v)} \qquad \text{(A.66)} \\
&= \frac{\det(1 - (\beta_v^\top\Sigma_v\beta_v)^{-1}\beta_v^\top\Sigma_v(\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1}\Sigma_v\beta_v)}{\det(1 - (\beta_v^\top\Sigma_v\beta_v)^{-1}\beta_v^\top\Sigma_v(\Sigma_v + \Sigma_\varepsilon)^{-1}\Sigma_v\beta_v)}. \qquad \text{(A.67)}
\end{aligned}$$

For the numerator, we write

$$\det(1 - (\beta_v^\top \Sigma_v \beta_v)^{-1} \beta_v^\top \Sigma_v (\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} \Sigma_v \beta_v) \tag{A.68}$$

$$= \det(1 - (n^{-2} \beta_v^\top \Sigma_v \beta_v)^{-1} (n^{-\frac{3}{2}} \beta_v^\top \Sigma_v) \tag{A.69}$$

$$\times (n^{-1}(\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon))^{-1} (n^{-\frac{3}{2}} \Sigma_v \beta_v)),$$

where each of the four terms premultiplied by a power of n has finite norm. The first term belongs to $\mathbb{R}^{k_v \times k_v}$ and satisfies

$$n^{-2} \beta_v^\top \Sigma_v \beta_v = n^{-2} \beta_v^\top (D_v + \beta_v \beta_v^\top) \beta_v \to B_v^{-2} \tag{A.70}$$

given that $\|D_v\|$ is bounded. From the second and fourth terms we obtain two types of terms involving the (bounded) diagonal matrix D_v. One of them is

$$n^{-\frac{3}{2}} \beta_v^\top D_v (n^{-1}(\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon))^{-1} (n^{-\frac{3}{2}} \beta_v \beta_v^\top \beta_v) \tag{A.71}$$

$$= n^{-\frac{1}{2}} \times [n^{-\frac{1}{2}} \beta_v^\top D_v][(\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} \beta_v][n^{-1} \beta_v^\top \beta_v],$$

where each term in square brackets is bounded. (For the second term, the argument is the same as the one leading to (A.59).) The other term,

$$n^{-\frac{3}{2}} \beta_v^\top D_v (n^{-1}(\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon))^{-1} (n^{-\frac{3}{2}} \beta_v D_v), \tag{A.72}$$

goes to zero obviously. The expression (A.68) consequently has the same limit as

$$\det(1 - n^{-2} B_v^{-2} \beta_v^\top \beta_v \beta_v^\top (\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} \beta_v \beta_v^\top \beta_v)$$

$$= \det(1 - \beta_v^\top (\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} \beta_v) \tag{A.73}$$

and similarly the denominator of (A.67) has the same limit as

$$\det(1 - \beta_v^\top (\Sigma_v + \Sigma_\varepsilon)^{-1} \beta_v) > 0, \tag{A.74}$$

which shows that η^{β_v} has the same limit as η^{F_v} given by (A.53), and therefore by (A.55). The inequality (A.74) holds because, by assumption, $\beta_v = \beta_\varepsilon a$, which implies

$$\begin{aligned}
&\det(1 - \beta_v^\top (\Sigma_v + \Sigma_\varepsilon)^{-1} \beta_v) \\
&\quad = \det(1 - a^\top \beta_\varepsilon^\top (D_v + D_\varepsilon + \beta_\varepsilon (1 + aa^\top) \beta_\varepsilon^\top)^{-1} \beta_\varepsilon a) \\
&\quad \geq \det(1 - \beta_\varepsilon a a^\top \beta_\varepsilon^\top (d + \beta_\varepsilon (1 + aa^\top) \beta_\varepsilon^\top)^{-1}) \\
&\quad = \det([n^{-1}(d + \beta_\varepsilon \beta_\varepsilon^\top)][n^{-1}(d + \beta_\varepsilon (1 + aa^\top) \beta_\varepsilon^\top)]^{-1}) \\
&\quad = \frac{\det(n^{-1}(d + \beta_\varepsilon^\top \beta_\varepsilon))}{\det(n^{-1}(d + \beta_\varepsilon^\top \beta_\varepsilon (1 + aa^\top)))} \\
&\quad > 0.
\end{aligned} \tag{A.75}$$

Therefore, we conclude that the total market inefficiency, in the limit, equals that of the set of portfolios $\{\beta_v\}$. To conclude that $\eta^{\beta_v}/\eta \to 1$, though, we need that $\eta > 0$ in the limit. From Equation (A.51), we need

$$\begin{aligned}&\det\left(\Sigma_\varepsilon^{-1}(\Sigma_\varepsilon + \bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)\right)\\&\quad\geq (1+\delta)\det\left((\Sigma_v+\Sigma_\varepsilon)^{-1}(\Sigma_v+\Sigma_\varepsilon+\bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)\right)\end{aligned} \tag{A.76}$$

for some $\delta > 0$. Rewriting the expressions inside the determinants, we seek

$$\det\left(1+\bar{\gamma}^2\Sigma_q^{\frac{1}{2}}\Sigma_\varepsilon\Sigma_q^{\frac{1}{2}}\right) \geq (1+\delta)\det\left(1+\bar{\gamma}^2\Sigma_q^{\frac{1}{2}}\Sigma_\varepsilon(\Sigma_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\Sigma_q^{\frac{1}{2}}\right).$$

We argued above—after Equation (A.62)—that

$$\det\left(1+\bar{\gamma}^2\Sigma_q^{\frac{1}{2}}\Sigma_\varepsilon\Sigma_q^{\frac{1}{2}}\right) \to \det\left(1+\bar{\gamma}^2\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top\beta_q\right). \tag{A.77}$$

Similarly, we have

$$\begin{aligned}&\det\left(1+\bar{\gamma}^2\Sigma_q^{\frac{1}{2}}\Sigma_\varepsilon(\Sigma_v+\Sigma_\varepsilon)^{-1}\Sigma_\varepsilon\Sigma_q^{\frac{1}{2}}\right)\\&\quad=\det\left(1+\bar{\gamma}^2\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top(\Sigma_v+\Sigma_\varepsilon)^{-1}\beta_\varepsilon\beta_\varepsilon^\top\beta_q\right)\\&\quad=\det\left(1+\bar{\gamma}^2\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top(1+\beta_v\beta_v^\top+\beta_\varepsilon\beta_\varepsilon^\top)^{-1}\beta_\varepsilon\beta_\varepsilon^\top\beta_q\right)\end{aligned} \tag{A.78}$$

in the limit; that is, the only contributors to the determinant are the low-rank matrices, with order-n eigenvalues. Using $\beta_v = \beta_\varepsilon a$ for the appropriate a and the push-through lemma, we write the last determinant above as

$$\begin{aligned}&\det\left(1+\bar{\gamma}^2\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top(1+\beta_v\beta_v^\top+\beta_\varepsilon\beta_\varepsilon^\top)^{-1}\beta_\varepsilon\beta_\varepsilon^\top\beta_q\right)\\&\quad=\det\left(1+\bar{\gamma}^2\beta_q^\top\beta_\varepsilon\beta_\varepsilon^\top(1+\beta_\varepsilon(1+aa^\top)\beta_\varepsilon^\top)^{-1}\beta_\varepsilon\beta_\varepsilon^\top\beta_q\right)\\&\quad=\det\left(1+\bar{\gamma}^2\beta_q^\top\beta_\varepsilon(1+\beta_\varepsilon^\top\beta_\varepsilon(1+aa^\top))^{-1}\beta_\varepsilon^\top\beta_\varepsilon\beta_\varepsilon^\top\beta_q\right)\\&\quad=\det\left(1+\bar{\gamma}^2B_{q\varepsilon}(n^{-1}+B_{\varepsilon\varepsilon}(1+aa^\top))^{-1}B_{\varepsilon\varepsilon}B_{\varepsilon q}\right)\\&\quad=\det\left(1+\bar{\gamma}^2B_{q\varepsilon}(n^{-1}+B_{\varepsilon\varepsilon}(1+aa^\top))^{-1}B_{\varepsilon\varepsilon}B_{\varepsilon q}\right)\\&\quad=\det\left(1+\bar{\gamma}^2B_{q\varepsilon}(n^{-1}B_{\varepsilon\varepsilon}^{-1}+(1+aa^\top))^{-1}B_{\varepsilon q}\right)\\&\quad\to\det\left(1+\bar{\gamma}^2B_{q\varepsilon}(1+aa^\top)^{-1}B_{\varepsilon q}\right).\end{aligned} \tag{A.79}$$

We continue by noting that

$$\begin{aligned}&\det\left(1+\bar{\gamma}^2B_{q\varepsilon}(1+aa^\top)^{-1}B_{\varepsilon q}\right)\\&\quad=\det\left(1+\bar{\gamma}^2B_{q\varepsilon}B_{\varepsilon q}-\bar{\gamma}^2B_{q\varepsilon}a(1+a^\top a)^{-1}a^\top B_{\varepsilon q}\right)\end{aligned} \tag{A.80}$$

with

$$\bar{\gamma}^2 B_{q\varepsilon} a (1 + a^\top a)^{-1} a^\top B_{\varepsilon q} > 0 \tag{A.81}$$

because $B_{q\varepsilon} a = \beta_q^\top \beta_v > 0$. The desired relation between expressions (A.77) and (A.80) is equivalent to

$$1 > \det\left(1 - \bar{\gamma}^2 (1 + \bar{\gamma}^2 B_{q\varepsilon} B_{\varepsilon q})^{-\frac{1}{2}} B_{q\varepsilon} a (1 + a^\top a)^{-1} a^\top B_{\varepsilon q} (1 + \bar{\gamma}^2 B_{q\varepsilon} B_{\varepsilon q})^{-\frac{1}{2}}\right),$$

which holds because the matrix on the right-hand side has eigenvalues lower than one, of which at least one strictly so. We now go further and compute the limit inefficiency of any set of portfolios ζ, which we normalize so that $n^{-1}\zeta^\top\zeta$ has a nonzero limit. The inefficiency equals

$$\begin{aligned}\eta^\zeta &= \frac{\det(\zeta^\top \mathrm{var}(v|\bar{p})\zeta)}{\det(\zeta^\top \mathrm{var}(v|\bar{s})\zeta)} \\ &= \frac{\det(\zeta^\top \Sigma_v \zeta - \zeta^\top \Sigma_v (\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} \Sigma_v \zeta)}{\det(\zeta^\top \Sigma_v \zeta - \zeta^\top \Sigma_v (\Sigma_v + \Sigma_\varepsilon)^{-1} \Sigma_v \zeta)},\end{aligned} \tag{A.82}$$

which we analyze using the same type of arguments as we used for (A.68). First, we project ζ on the columns of β_v, that is, write

$$\zeta = \beta_v a + \zeta_0 \equiv \zeta_1 + \zeta_0 \tag{A.83}$$

with a a matrix of appropriate dimension and $\zeta^\top \beta_v = 0$. In particular, the norms of all the terms of the matrices $n^{-2}\zeta^\top \Sigma_v \zeta$, $n^{-2}\zeta^\top \Sigma_v (\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2 \Sigma_\varepsilon \Sigma_q \Sigma_\varepsilon)^{-1} \Sigma_v \zeta)$, and $n^{-2}\zeta^\top \Sigma_v (\Sigma_v + \Sigma_\varepsilon)^{-1} \Sigma_v \zeta$ that contain ζ_0—and therefore $D_v \zeta_0$ rather than $\Sigma_v \zeta_0$—tend to zero as n goes to infinity. For instance,

$$\begin{aligned}&n^{-2} \zeta_1^\top \beta_v \beta_v^\top (\Sigma_v + \Sigma_\varepsilon)^{-1} D_v \zeta_0 \\ &\qquad = n^{-\frac{1}{2}} a^\top [n^{-1} \beta_v^\top \beta_v][\beta_v^\top (\Sigma_v + \Sigma_\varepsilon)^{-1}][n^{-\frac{1}{2}} D_v \zeta_0],\end{aligned} \tag{A.84}$$

with all the bracketed terms bounded in norm. Consequently, as long as the limit of $\det(n^{-2} a^\top (\beta_v^\top \Sigma_v \beta_v - \beta_v^\top \Sigma_v (\Sigma_v + \Sigma_\varepsilon)^{-1} \Sigma_v \beta_v) a)$ is strictly positive, that is, a is full rank, then $\eta^\zeta - \eta \to 0$.

The last case we consider is $a = 0$: a purely idiosyncratic portfolio. In this case, we scale the matrices in the numerator and denominator of (A.82) by n^{-1}. For the denominator, we have

$$\begin{aligned}&n^{-1}\zeta^\top D_v \zeta - n^{-1}\zeta^\top D_v (\Sigma_v + \Sigma_\varepsilon)^{-1} D_v \zeta \\ &\quad \geq n^{-1}\zeta^\top D_v \zeta - n^{-1}\zeta^\top D_v (D_v + D_\varepsilon)^{-1} D_v \zeta \\ &\quad > 0\end{aligned} \tag{A.85}$$

in the limit as long as Σ_ε is bounded below away from zero. We finally show that the difference between the matrices in the numerator and the denominator,

$$n^{-1}\zeta^\top D_v((\Sigma_v+\Sigma_\varepsilon)^{-1}-(\Sigma_v+\Sigma_\varepsilon+\bar\gamma^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1})D_v\zeta, \tag{A.86}$$

tends to zero by showing that

$$\|(\Sigma_v+\Sigma_\varepsilon)^{-1}-(\Sigma_v+\Sigma_\varepsilon+\bar\gamma^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1}\|\to 0. \tag{A.87}$$

Consider any portfolio u with $\|u\|=1$. With $D:=D_v+D_\varepsilon$, $\hat u=D^{-\frac12}u$, $\hat\beta_\varepsilon=D^{-\frac12}\beta_\varepsilon$, we decompose $\hat u$ into its projection on the columns of $\hat\beta_\varepsilon$ and an orthogonal component:

$$\hat u=\hat\beta_\varepsilon\kappa+\hat u_0 \tag{A.88}$$

with $\hat\beta_\varepsilon^\top\hat u_0=0$. We compute

$$\begin{aligned}
&u^\top(\Sigma_v+\Sigma_\varepsilon)^{-1}u\\
&\quad=\hat u^\top(1+\hat\beta_\varepsilon(1+\kappa\kappa^\top)\hat\beta_\varepsilon^\top)^{-1}\hat u\\
&\quad=\kappa^\top\hat\beta_\varepsilon^\top(1+\hat\beta_\varepsilon(1+\kappa\kappa^\top)\hat\beta_\varepsilon^\top)^{-1}\hat\beta_\varepsilon\kappa+\hat u_0^\top(1+\hat\beta_\varepsilon(1+\kappa\kappa^\top)\hat\beta_\varepsilon^\top)^{-1}\hat u_0\\
&\quad\to\hat u_0^\top\hat u_0,
\end{aligned} \tag{A.89}$$

because the first term is bounded above by $\kappa^\top\kappa$ and therefore tends to zero. We also have

$$\begin{aligned}
&u^\top(\Sigma_v+\Sigma_\varepsilon+\bar\gamma^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1}u\\
&\qquad\geq\hat u^\top(1+\hat\beta_\varepsilon(1+\kappa\kappa^\top)\hat\beta_\varepsilon^\top+n^{-1}cD^{-\frac12}\Sigma_\varepsilon\Sigma_\varepsilon D^{-\frac12})^{-1}\hat u
\end{aligned} \tag{A.90}$$

Let $L=1+\kappa\kappa^\top\geq 1$ to simplify notation. Decompose the last term in the equation above by writing $\Sigma_\varepsilon=D_\varepsilon+\beta_\varepsilon\beta_\varepsilon^\top$ with the goal of computing

$$\hat u_0^\top(1+\hat\beta_\varepsilon L\hat\beta_\varepsilon^\top+n^{-1}cD^{-1}D_\varepsilon^2+c\hat\beta_\varepsilon^\top B_\varepsilon\hat\beta_\varepsilon+n^{-1}cD_\varepsilon\hat\beta_\varepsilon\hat\beta_\varepsilon^\top+n^{-1}c\hat\beta_\varepsilon\hat\beta_\varepsilon^\top D_\varepsilon)^{-1}\hat u_0.$$

Let

$$M\equiv(1+\hat\beta_\varepsilon(L+cB_\varepsilon)\hat\beta_\varepsilon^\top)^{-\frac12}. \tag{A.91}$$

By using bounding arguments and combining terms, we can focus on

$$\begin{aligned}
&\hat u_0^\top(1+\hat\beta_\varepsilon(L+cB_\varepsilon)\hat\beta_\varepsilon^\top+n^{-1}cD_\varepsilon\hat\beta_\varepsilon\hat\beta_\varepsilon^\top+n^{-1}c\hat\beta_\varepsilon\hat\beta_\varepsilon^\top D_\varepsilon)^{-1}\hat u_0 &\text{(A.92)}\\
&\quad=\hat u_0^\top M(1+n^{-1}cM(D_\varepsilon\hat\beta_\varepsilon\hat\beta_\varepsilon^\top+\hat\beta_\varepsilon\hat\beta_\varepsilon^\top D_\varepsilon)M)^{-1}M\hat u_0\\
&\quad=\hat u_0^\top(1+n^{-1}cM(D_\varepsilon\hat\beta_\varepsilon\hat\beta_\varepsilon^\top+\hat\beta_\varepsilon\hat\beta_\varepsilon^\top D_\varepsilon)M)^{-1}\hat u_0.\\
&\quad=:\hat u_0^\top(1+N)^{-1}\hat u_0. &\text{(A.93)}
\end{aligned}$$

We note that

$$\begin{aligned}
&\|n^{-1}cM\hat{\beta}_\varepsilon\hat{\beta}_\varepsilon^\top D_\varepsilon M\hat{u}_0\|^2 \\
&\quad = n^{-2}c^2\hat{u}_0^\top M D_\varepsilon\hat{\beta}_\varepsilon\hat{\beta}_\varepsilon^\top M^2\hat{\beta}_\varepsilon\hat{\beta}_\varepsilon^\top D_\varepsilon M\hat{u}_0 \\
&\quad \geq n^{-2}c^2\hat{u}_0^\top M D_\varepsilon\hat{\beta}_\varepsilon\hat{\beta}_\varepsilon^\top D_\varepsilon M\hat{u}_0 \\
&\quad = n^{-2}c^2\hat{u}_0^\top D_\varepsilon\hat{\beta}_\varepsilon\hat{\beta}_\varepsilon^\top D_\varepsilon\hat{u}_0 \\
&\quad \sim n^{-1},
\end{aligned} \tag{A.94}$$

so that $N\hat{u}_0 \sim n^{-\frac{1}{2}}$. Consequently,

$$\hat{u}_0^\top\hat{u}_0 - \hat{u}_0^\top(1+N)^{-1}\hat{u}_0 = \hat{u}_0^\top(1+N)^{-1}N\hat{u}_0 \sim n^{-\frac{1}{2}}$$

given that $1+N \geq 1$. It is clear that, for $\hat{u}_1 = \hat{\beta}_\varepsilon\kappa$, we have

$$\hat{u}_1^\top(\Sigma_v + \Sigma_\varepsilon + \bar{\gamma}^2\Sigma_\varepsilon\Sigma_q\Sigma_\varepsilon)^{-1}\hat{u}_1 \leq \hat{u}_1^\top(\Sigma_v+\Sigma_\varepsilon)^{-1}\hat{u}_1 \to 0,$$

so that the two inverse matrices in Equation (A.87) tend to each other in norm on each of two orthogonal components of $\mathbb{R}^n$. This step completes the proof of the last statement of the proposition. ■ Let $\beta_i \in \mathbb{R}^{n\times k_i}$, $i \in \{1,2\}$, be such that $n^{-1}\beta_i^\top\beta_i$ has a well-defined strictly positive limit and $M \in \mathbb{R}^{n\times n}$ symmetric, $0 < \underline{m} < \overline{m} \in \mathbb{R}$ s.t. $\underline{m} < M < \overline{m}$. Then

$$\left(M + \beta_1\beta_1^\top + \beta_2\beta_2^\top\right)^{-1}\beta_1 \sim n^{-\frac{1}{2}} \tag{A.95}$$

and consequently the matrix

$$\left(M + \beta_1\beta_1^\top + \beta_2\beta_2^\top\right)^{-1}\beta_1\beta_1^\top \tag{A.96}$$

is uniformly bounded in norm.

Proof of Lemma A.5. We show that

$$\|\left(M + \beta_1\beta_1^\top + \beta_2\beta_2^\top\right)^{-1}\beta_1\| \sim n^{-\frac{1}{2}} \tag{A.97}$$

by showing that the same holds when replacing the last instance of β_1 with any of its columns. Consider the eigenvectors ζ of $\beta_1\beta_1^\top + \beta_2\beta_2^\top$ corresponding to nonzero eigenvalues. All these eigenvectors are generated by the columns of β_1 and β_2, and the eigenvalues are of order n. Let b be a column of β_1 and decompose it as $b = \sum_i a_i\zeta_i$. We show that

$$\|\left(M + \beta_1\beta_1^\top + \beta_2\beta_2^\top\right)^{-1}\zeta_i\| \sim n^{-1}. \tag{A.98}$$

Since

$$\left(\beta_1\beta_1^\top + \beta_2\beta_2^\top\right)\zeta_i = \lambda_i n\zeta_i, \tag{A.99}$$

it follows that

$$\left(M+\beta_1\beta_1^\top+\beta_2\beta_2^\top\right)^{-1}\zeta_i=(\lambda_i n)^{-1}\zeta_i-(\lambda_i n)^{-1}\left(M+\beta_1\beta_1^\top+\beta_2\beta_2^\top\right)^{-1}M\zeta_i, \quad \text{(A.100)}$$

which implies the desired conclusion given that ζ_i, $\left(M+\beta_1\beta_1^\top+\beta_2\beta_2^\top\right)^{-1}$, and M are bounded in norm. From the fact that $\|b\|^2=\sum_i a_i^2\sim n$ it follows that

$$\|\left(M+\beta_1\beta_1^\top+\beta_2\beta_2^\top\right)^{-1}b\|\sim n^{-\frac{1}{2}}, \quad \text{(A.101)}$$

as sought. ■ [Push-through lemma] Consider matrices V, U, and X of appropriate dimensions with $XU=0$. Then

$$(I+UV+X)^{-1}U=U(I+VU)^{-1}. \quad \text{(A.102)}$$

Proof of Lemma A.5. Start with

$$\begin{aligned}(I+UV+X)^{-1}U(I+VU)&=(I+UV+X)^{-1}(I+UV)U\\&=(I+UV+X)^{-1}(I+UV+X)U\\&=U\end{aligned}$$

and multiply both sides by $(I+VU)^{-1}$. ■

Proof of examples 1 and 2. The statements made in examples 1 and 2 are based on results derived formally in the paper, as follows. First, note that in all of these examples Assumption 1″ holds, in addition to Assumption 3. Example 1 can be viewed as a consequence of Propositions 4(d) and 5 and of Equation (A.29), which expresses the eigenvalues of $\Sigma_{v|p}^{-1}\Sigma_{v|s}$. Example 2 is a consequence of the observation that, as shown in Lemma A.5, the eigenvectors of $\Sigma_{v|p}^{-1}\Sigma_{v|s}$ include the eigenvectors of $O=\beta\beta^\top$, and these capture the entire inefficiency by Proposition 5.

Proof of Proposition 6. As noted in Section A.2, at any interior equilibrium, $g^I(I,M)=0$ and $g^M(I,M)=0$. The second of these equations defines implicitly the function

$$\mathcal{M}(I)=\frac{I}{k}\left(\frac{\eta(I)}{2\gamma}+\frac{k_p-k_a}{2}\right), \quad \text{(A.103)}$$

while setting the right-hand side of Equation (A.3) to zero defines a function $\mathcal{I}(M)$, given that c increases with I and η decreases with I. Further, since this right-hand side increases with I and decreases with M, $\mathcal{I}(M)$ is increasing. As we specified in the text, we concentrate on the equilibrium with the maximum I, which is assumed to be interior. Since $M<\bar{M}$, $g^M(\mathcal{I}(\bar{M}),\bar{M})>0$. On the other hand, by the definition of $\mathcal{I}$, $g^I(\mathcal{I}(\bar{M}),\bar{M})\leq 0$. Given that

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

$$\frac{d}{dM}\left(g^M(\mathcal{I}(M),M)-g^I(\mathcal{I}(M),M)\right)>0. \tag{A.104}$$

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

$$g^M_M-g^M_I\frac{g^I_M}{g^I_I}>0 \tag{A.105}$$

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

$$g^I_M g^M_I<g^I_I g^M_M. \tag{A.106}$$

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

$$\begin{pmatrix} g^M_I & g^M_M \\ g^I_I & g^I_M \end{pmatrix}\begin{pmatrix} I_k \\ M_k \end{pmatrix}=-\frac{M}{I}\begin{pmatrix}1\\0\end{pmatrix}, \tag{A.107}$$

and therefore by

$$\begin{pmatrix} I_k \\ M_k \end{pmatrix}=\frac{1}{g^I_M g^M_I-g^I_I g^M_M}\begin{pmatrix} g^I_M \\ -g^I_I \end{pmatrix}\left(-\frac{M}{I}\right). \tag{A.108}$$

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.106), we conclude $I_k<0$ and $M_k<0$. A decrease in k, therefore, translates into a higher I and into a higher M. The number of self-directed investors is clearly unaffected: f_a decreases from Equation (10) 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, that is, as a quadratic form—with I, because

$$\Sigma_{v|p}=\Sigma_v-\Sigma_v\left(\Sigma_v+\Sigma_\varepsilon+\theta_q\Sigma_q\theta_q\right)^{-1}\Sigma_v \tag{A.109}$$

and $\theta_q=\frac{\gamma}{I}\Sigma_\varepsilon$. (The one nonobvious 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. For values $k_1>k_2$ of k, equilibrium values $I_1(n)<I_2(n)$ obtain for every n. Denoting the limit points of these two series by I_1 and I_2, respectively, we have $I_1\leq I_2$, and in fact $I_1<I_2$ because the limit equilibrium (I,M), too, must change with k to satisfy $g^M=0$. At the same time, the corresponding differentials in inefficiency, $\eta_1^{\beta_v}(n)-\eta_2^{\beta_v}(n)$ and $\eta_1(n)-\eta_2(n)$, have the same limit with n: the macro-efficiency accounts for the entire change in efficiency.

Furthermore, this change is nonzero because the limit inefficiency sensitivity to I is. ■

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

$$\begin{pmatrix} g_I^M & g_M^M \\ g_I^I & g_M^I \end{pmatrix} \begin{pmatrix} I_{k_p} \\ M_{k_p} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{A.110}$$

and therefore by

$$\begin{pmatrix} I_{k_p} \\ M_{k_p} \end{pmatrix} = \frac{1}{2} \frac{1}{g_M^I g_I^M - g_I^I g_M^M} \begin{pmatrix} g_M^I - g_M^M \\ g_I^M - g_I^I \end{pmatrix}. \tag{A.111}$$

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.106). Thus, both I and M decrease as k_p decreases. Consequently, the inefficiency η increases. Following the same line of argument as in the proof of Proposition 6, the macro-inefficiency, which accounts for the entire inefficiency in the limit, also accounts for its entire sensitivity to the cost k_p. Since M decreases, while the expression $\frac{M}{I}k - c(M, I)$ remains equal to zero, I 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 8. (a) Consider demands conditional on realized supply:

$$\begin{aligned} \mathrm{E}[x_i|q] &= \left(\gamma\Sigma_{v|s}\right)^{-1} \left(\bar{v} - \mathrm{E}[p|q]\right) \\ &= \left(\gamma\Sigma_{v|s}\right)^{-1} \pi + \left(\gamma\Sigma_{v|s}\right)^{-1} \theta_s\theta_q(q - \bar{q}) \qquad \text{(A.112)} \\ \mathrm{E}[x_u|q] &= \left(\gamma\Sigma_{v|p}\right)^{-1} \left(\mathrm{E}[\mathrm{E}[v|p]|q] - \mathrm{E}[p|q]\right) \\ &= \left(\gamma\Sigma_{v|p}\right)^{-1} \pi + \left(\gamma\Sigma_{v|p}\right)^{-1} \left(\theta_s - \Sigma_v\Sigma_{\hat{p}}^{-1}\right) \theta_q(q - \bar{q}), \qquad \text{(A.113)} \end{aligned}$$

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.

$$\theta_s > \theta_s - \Sigma_v\Sigma_{\hat{p}}^{-1} \tag{A.114}$$

$$\theta_s\theta_q > \left(\theta_s - \Sigma_v\Sigma_{\hat{p}}^{-1}\right) \theta_q \tag{A.115}$$

$$\left(\gamma\Sigma_{v|s}\right)^{-1} \theta_s\theta_q > \left(\gamma\Sigma_{v|p}\right)^{-1} \left(\theta_s - \Sigma_v\Sigma_{\hat{p}}^{-1}\right) \theta_q. \tag{A.116}$$

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

$$\begin{aligned}
\mathrm{E}[x_i|s] &= \left(\gamma\Sigma_{v|s}\right)^{-1}\left(\mathrm{E}[v|s] - \mathrm{E}[p|s]\right) \\
&= \left(\gamma\Sigma_{v|s}\right)^{-1}\pi + \left(\gamma\Sigma_{v|s}\right)^{-1}\left(\Sigma_v\Sigma_s^{-1} - \theta_s\right)(s - \bar{v}) \qquad \text{(A.117)} \\
\mathrm{E}[x_u|s] &= \left(\gamma\Sigma_{v|p}\right)^{-1}\left(\mathrm{E}[\mathrm{E}[v|p]|s] - \mathrm{E}[p|s]\right) \\
&= \left(\gamma\Sigma_{v|p}\right)^{-1}\pi + \left(\gamma\Sigma_{v|p}\right)^{-1}\left(\Sigma_v\Sigma_{\hat{p}}^{-1} - \theta_s\right)(s - \bar{v}). \qquad \text{(A.118)}
\end{aligned}$$

Under Assumption 1 or Assumption 2, $\Sigma_v\Sigma_{\hat{p}}^{-1} < \theta_s < \Sigma_v\Sigma_s^{-1}$ (see Equation (A.13)). (In general, because of market clearing, $I\mathrm{E}[x_i|s] + U\mathrm{E}[x_u|s] = \bar{q}$.) The conclusion follows. ■

Proof of Proposition 9. The logic of the proof is the same as for Proposition 7. Specifically, we are solving for the derivatives I_z and M_z from

$$\begin{pmatrix} g_I^M & g_M^M \\ g_I^I & g_M^I \end{pmatrix}\begin{pmatrix} I_z \\ M_z \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\frac{\partial\eta}{\partial z} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \text{(A.119)}$$

giving

$$\begin{pmatrix} I_z \\ M_z \end{pmatrix} = \frac{1}{g_M^I g_I^M - g_I^I g_M^M}\begin{pmatrix} g_M^I - g_M^M \\ g_I^M - g_I^I \end{pmatrix}\begin{pmatrix} \frac{1}{2}\frac{\partial\eta}{\partial z} \end{pmatrix}. \qquad \text{(A.120)}$$

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, while S_p decreases because $I + S_p$ does not change. The total effect on the inefficiency η is ambiguous. ■