Why Do Demand Curves for Stocks Slope Down?

Antti Petajisto*

Yale School of Management

International Center for Finance

P.O. Box 208200

New Haven, CT 06520-8200

antti.petajisto@yale.edu

November 29, 2004

Abstract

Representative agent models including the CAPM are inconsistent with existing empirical evidence for steep demand curves for individual stocks. This paper resolves the issue by proposing that stock prices are instead set by two separate classes of investors. The market portfolio is priced by individual investors based on their collective risk aversion. Individual investors also delegate part of their wealth to active money managers who use that wealth to price stocks in the cross-section. In equilibrium the fee charged by active managers has to equal the before-fee alpha they earn. This produces wide pricing bounds for individual stocks and can account for several empirically observed puzzles such as the magnitude of the S&P 500 index premium.

JEL classification: G12, G14, G20, D50

Keywords: Demand curves for stocks; delegated portfolio management; equilibrium mispricing; index premium

*I wish to thank Steve Ross, Dimitri Vayanos, Tuomo Vuolteenaho, and Jiang Wang for valuable discussions, as well as Nick Barberis, Randy Cohen, Josh Coval, John Griffin, Jonathan Ingersoll, Jonathan Lewellen, Andrew Lo, Robert Merton, Oguzhan Ozbas, Anna Pavlova, Andrei Shleifer, Matthew Spiegel, Jeremy Stein, Ivo Welch and seminar participants at Harvard University, Helsinki School of Economics, INSEAD, Massachusetts Institute of Technology, Rice University, University of Alberta, University of Michigan, Yale University, and EFMA 2003 annual meeting for comments. I am also grateful to the Emil Aaltonen Foundation and the Finnish Cultural Foundation for financial support.

1 Introduction

On July 9, 2002, Standard and Poor's announced that it would delete all seven non-U.S. firms from its S&P 500 index and replace them with U.S. firms. The changes were to take place on July 19, and they included large firms like Royal Dutch, Unilever, Goldman Sachs, and UPS. The day following the announcement the deleted firms fell by an average of 3.7% while the added firms went up by 5.9% relative to the value-weighted market index, reportedly on trading by hedge funds and active managers.[1] During the ten days leading to the effective day the cumulative market-adjusted return was -6.6% for the deletions and $+12.3\%$ for the additions – all on a bureaucratic event which contained absolutely no news about the level or riskiness of the cash flows of the firms involved. In spite of its size and publicity, this event produced a very significant price impact which showed no signs of reversal, at least in the following two months.[2]

To explain this kind of price effects, a growing literature has proposed that demand curves for stocks slope down (Shleifer (1986) is an early reference). Whenever stocks are added to or deleted from a popular stock market index, index funds buy the additions and sell the deletions. For the S&P 500, mechanical indexers currently hold about 10% of the market value of every stock in the index. When the demand curve for a stock has a nonzero slope, the large supply shocks due to indexers can move prices. The typical price effect for both additions and deletions has been about 15% for the S&P 500 in 2000, and other widely tracked indices have exhibited comparable demand elasticities (Petajisto (2004)).

However, this explanation presents a puzzle: How do we reconcile the large magnitude of the price effect with asset pricing theory? In neoclassical finance, price equals expected future cash flows discounted by systematic risk, so the demand curve for a stock should be (almost) perfectly horizontal and we should observe (virtually) no price impact. Asymmetric information models[3] cannot explain the significant price effects, because the puzzle here has to do with *clearly uninformed* supply shocks as illustrated by the above S&P 500 event.[4]

The limits of arbitrage literature has been suggested as a way to bridge the gap be-

[1] The Wall Street Journal, 7/11/02.

[2] For illustration, see Figure 3 in the appendix.

[3] Some examples are Grossman and Stiglitz (1980), Glosten and Milgrom (1985), and Kyle (1985).

[4] Denis, McConnell, Ovtchinnikov, and Yu (2003) actually find evidence that regular S&P 500 index

tween theory and empirical work (Barberis and Thaler (2003) and Wurgler and Zhuravskaya (2002)). Mechanisms such as noise trader risk (De Long et al. (1990)) and performance-based arbitrage (Shleifer and Vishny (1997)) can indeed influence the pricing of non-diversifiable risk, but they cannot explain why investors are so reluctant to take *diversified* positions in individual stocks.

This paper first shows that existing equilibrium models underestimate the actual slopes of demand curves for stocks by several orders of magnitude. It then proposes a theoretical equilibrium model that can produce a realistically large magnitude for the slopes of demand curves, and not only for index additions and deletions but for *all* the stocks in the economy.

Despite the frequent references to indexing, this paper is about much more than that – indexing just happens to provide a relatively clean empirical test. Downward-sloping demand curves have the broad implication that assets need not be priced exactly at their fundamental values. Even completely uninformed supply shocks can move prices. Hence, our explanation for downward-sloping demand curves will also serve as a theory of approximate pricing of assets in equilibrium.

To illustrate the failure of traditional pricing models, consider the following CAPM calibration: The U.S. stock market capitalization at the end of 2002 was about $11 trillion, which means that collectively people invested $11 trillion in the market portfolio, perhaps expecting about a 5% annual risk premium and 20% annual volatility. This information allows us to back out their risk aversion. Now let us assume that the price of one stock changes slightly for noninformational reasons so that the investors suddenly perceive the stock to have an annual alpha of +1%, with idiosyncratic annual volatility of 30%. The investors should then immediately pour $1 trillion into that stock – more than three times the market capitalization of General Electric.[5] In other words, even a 1% annual alpha would be absurdly large in a CAPM setting. A representative investor who is willing to invest $11 trillion in the market portfolio should be extremely aggressive when any mispricings occur for individual stocks. More generally, this calibration shows that no

changes (unlike the event we picked) may not be completely free of information. But index changes even for mechanical rule-based indices such as the Russell 2000 exhibit comparable price effects.

[5] The optimal dollar investment for a CARA or CRRA investor is proportional to $\frac{\mu}{\sigma^2}$. This is $\frac{0.05}{0.2^2} = 1.25$ for the market portfolio and $\frac{0.01}{0.3^2} = 0.11$ for the idiosyncratic gamble, producing a dollar investment of $\frac{0.11}{1.25} \times \$11 = \$1$ trillion in the idiosyncratic gamble.

model with a single representative investor can simultaneously generate realistic demand curves for individual stocks and a plausible market risk premium.

Building on this key insight, the story in our paper is that demand curves seem too steep only when we assume the *same* group of investors prices both the market portfolio and the cross-section of individual stocks. The puzzle disappears if instead individual investors delegate the picking of individual stocks to professional active managers who charge a fee for assets under management. The slope of the demand curves for individual stocks is then set by the aggressiveness of the active managers, but since the active managers simply invest all of the wealth under their management, the demand curve slopes are really determined by how much money the individual investors allocate to the active managers. The less money the active managers have, the less aggressive they are, the steeper the demand curves, the greater the relative mispricings for individual stocks, and the greater the alpha earned by the active managers. In equilibrium, individual investors should be indifferent between the market portfolio and an actively managed portfolio, so the active managers' after-fee alpha should be zero. This means that equilibrium demand curves for stocks will be just steep enough to let active managers earn a before-fee alpha that covers their fee.

We formalize our story in a simple model similar to the CAPM setting. There are only two differences: First, we assume there is a fixed cost for actively managing a stock portfolio; if one does not pay the cost, one can only invest in the market portfolio. We interpret this as costly information acquisition; if one does not know about individual stocks, one's best bet is the market portfolio. Second, we assume the fixed cost is paid through a financial institution as a proportional fee.

The model thus introduces a layer of financial institutions between stocks and end investors. Active managers act as stock pickers, using all their delegated wealth to take positions in individual stocks, and they charge a fee for their services. End investors then choose their optimal allocation of wealth between an actively managed portfolio, a passively managed market portfolio (with zero fee), and a risk-free asset. The remaining supply of each stock is passively held by exogenous noise traders – without this group, even the active managers would just have to hold the market portfolio. We do not consider agency issues, so the only friction we introduce relative to the CAPM is the fixed cost which generates a fee for active management.

We find that this delegation of portfolio management completely changes the cross-sectional pricing of stocks. Now the slopes of demand curves are no longer determined by end investors' risk aversion – instead they depend on the wealth allocated to active managers, which in turn depends on the fee charged by the active managers. If the fee is 1.5% per year, then the typical stock will be "mispriced" so that it will have an alpha of either +1.5% or −1.5% per year in equilibrium. If on average such mispricings are corrected slowly over several years,[6] then these annual alphas will be capitalized into much greater variation in stock prices today. E.g. an annual alpha of +1%, fully corrected over 5 years, means a stock is underpriced by 5%. Thus the initial mispricings created by the management fee are further magnified by their slow expected convergence to fundamental values, and this allows economically large fluctuation in stock prices today.[7] For comparison, if we set the active managers' fee to zero, pricing collapses to the traditional CAPM benchmark where annual alphas are always well within 1 bp from zero.

Yet the presence of institutions does not create any friction in the model – the true source of friction is the underlying fixed cost. In fact, we show that institutions actually mitigate the effect of the fixed cost and produce the *flattest* possible demand curves. This is because institutions allow the risk of the actively managed portfolio to be shared among all investors in the economy, while without institutions it is not possible to induce such complete risk sharing in equilibrium as investors would then drive alphas too close to zero and thus not be willing to pay the participation cost. Hence, the underlying fixed cost gives rise to endogenous institutions which make stock prices more efficient, i.e. closer to the CAPM benchmark. Consistent with the predictions of functional and structural finance (Merton and Bodie (2002)), our institutional structure minimizes price distortions due to the underlying market friction.

Empirical evidence appears generally consistent with our equilibrium. Wermers (2000) and Daniel et al. (1997) find that active fund managers do have stock-selection ability, especially if they concentrate on relatively few industries (Kacperczyk, Sialm, and Zheng (2004)) or if they are small (Chen, Hong, Huang, and Kubik (2004)), but once their fees

[6] For empirical evidence on the slow correction of mispricings, see e.g. Cohen, Gompers, and Vuolteenaho (2002).

[7] The stochastic discount factor of the economy consists of a market term as well as a number of ad hoc terms corresponding to idiosyncratic mispricings. Aside from market risk, there is no economic structure to it. This is precisely what we would expect – if all the mispricings were easily identifiable by all investors, there would be no reason for anyone to conduct fundamental analysis of individual stocks.

and expenses are taken into account, their alphas fall back to approximately zero. For our pricing results it is crucial that active managers indeed earn positive before-fee alphas, but whether their alphas exactly cover their fees does not matter that much.

Our paper is certainly not the first one in the theoretical finance literature to have steep demand curves – many usually single-asset models exogenously assume them.[8] This assumption has allowed them to describe interesting rational as well as behavioral price effects for individual stocks, and it is amply justified by empirical evidence, but it ignores the contradiction with the neoclassical multi-asset benchmarks of CAPM and APT. In contrast, the sole purpose of our model is to produce such steep demand curves as an endogenous equilibrium outcome.

Multi-asset equilibrium models such as Admati (1985) and Merton (1987) face the same problem as the CAPM. Whenever the cross-sectional pricing of stocks is determined by the same investors who collectively hold the entire market portfolio, uninformed supply shocks can no longer move alphas by more than a negligible amount, so demand curves will have to be horizontal.[9]

Our model may resemble a multi-asset generalization of information cost or participation cost models including Grossman and Stiglitz (1980), Grossman and Miller (1988), and Allen and Gale (1994). However, rather than individuals paying the cost themselves, we explicitly introduce intermediaries that pay the cost, compete for other investors' funds, and then charge a proportional fee to their investors. The distinction is crucial for two reasons: First, an investor who has paid a fixed cost has an enormous incentive to act as an intermediary for other investors, and if he does that, his investment aggressiveness is no longer determined by his personal risk aversion. Second, it allows us to calibrate the model to observable parameter values (namely, the percentage fee) and still obtain dramatic pricing

[8] This covers virtually all single-asset models where agents are not risk-neutral and thus their risk aversion plays a role in pricing (e.g. Chen, Hong, and Stein (2002), Allen and Gale (1994), and many others). Also some multi-asset models (e.g. Barberis and Shleifer (2002) and Wurgler and Zhuravskaya (2002)) exogenously assume steep demand curves.

[9] Hence, exogenous tastes for individual stocks as in Fama and French (2004) can only produce negligible deviations from CAPM pricing. Gomes, Kogan, and Zhang (2003) offer an example of a multi-asset equilibrium where interesting price effects emerge from a conditional CAPM but where demand curves are still horizontal. In Daniel, Hirshleifer, and Subrahmanyam (2001), systematic risk can be mispriced but again individual stocks cannot meaningfully deviate from factor pricing.

results. Our model also shares resemblance to Berk and Green (2004) where in equilibrium active funds have to earn their fees, but their paper focuses on the dynamics of the mutual fund industry while our paper concentrates on equilibrium prices of stocks in the presence of active mutual funds.

Besides steep demand curves, some empirical papers have suggested altogether different mechanisms for generating the observed price effects around index changes. The most prominent include liquidity (e.g. Edmister, Graham, and Pirie (1994) and Chordia (2001)), information (e.g. Denis et al. (2003)), and market segmentation (e.g. Chen, Noronha, and Singal (2004)). While these hypotheses could well contribute to the effect, each of them tends to be specific to a particular type of event. Given all the diverse contexts where we observe similar price effects, one might naturally look for a more general explanation that can produce steep demand curves across the board. Moreover, since these alternatives have not been formalized in this context, we do not know how large effects they can theoretically explain.

Our paper makes two main contributions. First, it presents the first generally applicable explanation for downward-sloping demand curves which gets the magnitude of the effect approximately right. Thus it provides a theoretical justification for the models that have exogenously assumed steep demand curves. Second, it illustrates that financial institutions do indeed matter for asset pricing. This is in contrast to all models based on a single representative agent, suggesting that such models may be better suited for pricing systematic risk than a wide cross-section of stocks with idiosyncratic risk. Furthermore, we obtain our result entirely without agency issues, complementing the existing literature (e.g. Ross (1989) and Allen (2001)) which has pointed out the relevance of institutions to asset pricing due to agency issues.

The paper proceeds as follows. Section 2 starts with a simple CAPM benchmark and contrasts it with empirical evidence to illustrate the puzzle. It also briefly addresses alternative hypotheses in the literature. Section 3 presents our model and the equilibrium, and it provides a numerical calibration to show the magnitudes of the predicted effects. Section 4 presents the other empirical predictions of the model. Section 5 discusses interpretations and extensions of the model, and section 6 concludes. The appendix contains all algebra as well as most figures.

2 The Puzzle: Theory and Empirical Evidence

No equilibrium model of course literally implies that the demand curve for a stock is perfectly horizontal.[10] The real question here is about the magnitude of the slope: Is it really "negligible" as suggested by the neoclassical models, or does it deviate "significantly" from zero? In other words, can we assume for practical purposes that the stock price is unaffected by the supply of the stock? We start by presenting a simple CAPM calibration to see what exactly a negligible price impact would mean.

2.1 A Simple CAPM Calibration

Let there be N_S stocks with a supply of 1 unit each, and a risk-free asset with an infinitely elastic supply. One period from now stock i pays a liquidating dividend of $\widetilde{x}_i = a_i + b_i\widetilde{y} + \widetilde{e}_i$. Systematic shocks to the economy are represented by the unexpected return on the market portfolio $\widetilde{y} \sim N\left(0, \sigma_m^2\right)$, idiosyncratic shocks to the stock are denoted by $\widetilde{e}_i \sim N\left(0, \sigma_{e_i}^2\right)$, and a_i and b_i are stock-specific constants.[11] The return on the risk-free asset is normalized to zero.

The economy is populated by mean-variance investors who can be aggregated into a representative investor with CARA utility and a coefficient of absolute risk aversion γ.

The representative investor's maximization problem is:

$$\max_{\{\theta_i\}} E\left[-\exp\left(-\gamma\widetilde{W}\right)\right]$$

$$\text{s.t.} \quad \widetilde{W} = W_0 + \sum_{i=1}^{N_S} \theta_i \left(\widetilde{x}_i - P_i\right). \qquad (1)$$

We calculate the first-order conditions with respect to θ_i, taking the market variance σ_m^2 as exogenous. We denote the equilibrium supply held by the investor as u_i, and we plug it in

[10] When the representative investor buys more of a stock, that stock becomes a larger part of his systematic portfolio risk, i.e. its beta increases, and thus it requires a higher return. However, in a well-diversified portfolio, the stock should represent only a tiny fraction of the portfolio anyway, so this effect should be negligible.

[11] Since the market return is a value-weighted return on individual stocks, the idiosyncratic stock returns have to add up to zero. We ignore this constraint for analytical convenience. This has a negligible impact on our results when there is a large number of assets.

for θ_i. This gives us the equilibrium price:

$$P_i = a_i - \gamma \left[\underbrace{\sigma_m^2 \left(\sum_{j \neq i} u_j b_j \right) b_i}_{\text{depends on systematic risk } b_i} + \underbrace{\left(\sigma_m^2 b_i^2 + \sigma_{e_i}^2 \right) u_i}_{\text{depends on supply } u_i} \right]. \tag{2}$$

The price is equal to the expected payoff a_i minus a discount, where the price discount will be dominated by the term that does not depend on the stock's supply.

We pick a one-year holding period, $N_S = 1,000$, $a_i = 105$, $b_i = 100$, and $\sigma_{e_i}^2 = 900$ for all stocks and $\sigma_m^2 = 0.04$ for the market variance. We start by letting the representative investor hold the entire market portfolio, so that $u_i = 1$ for all stocks. We also set $\gamma = 1.247 \times 10^{-5}$ which produces an equilibrium market risk premium of 5%. Each stock will then have a price of 100, market beta of 1, and idiosyncratic standard deviation of return of 30%.

Now consider a supply shock of -10% to a stock. Suppose, for example, that a new investor enters the market and buys 10% of the shares of stock i. Plugging in $u_i = 0.9$, the price of stock i will then increase to 100.00162. In other words, this supply shock will produce a 0.16 basis point price impact. Part of this impact is due to the decreased supply of market risk and in fact all stocks would go up by 0.05 bp for this reason, so relative to the other stocks this stock would go up by 0.11 bp. This is what the "almost perfectly horizontal" demand curves mean.[12]

What is the intuition for the result? In equilibrium, the representative investor is willing to bear a large amount of systematic market risk for a risk premium of 5%. Given that he holds large number of stocks (1,000), a 10% supply shock to an individual stock is only a tiny fraction of his entire portfolio (1/10,000). If he requires a 5% risk premium for an investment equal to the size of his entire portfolio, he will need only a tiny fraction of that premium for an investment equal to a tiny fraction of his entire portfolio.

2.2 Empirical Evidence

To estimate the slope of the demand curve for a stock, most studies focus on large supply shocks where the source can be identified as uninformed both by market participants and

[12] These results are not affected by the choice of CARA utility as opposed to CRRA utility. See section 5.4 for more details on CRRA utility.

the econometrician. One possible sample is provided by large block trades, studied by e.g. Scholes (1972) and Holthausen and Leftwich (1987). Seasoned equity offerings provide another experiment, studied by e.g. Loderer, Cooney, and van Drunen (1991). Except for the early study by Scholes, these papers typically find relatively small negative values for the price elasticity of demand (e.g. a median of -4.31 and mean of -11.1 for Loderer et al.).[13] Trading due to merger arbitrage strategies also seems to produce a significant price impact (Mitchell, Pulvino, and Stafford (2004)) and could be used to extract elasticity estimates. Nevertheless, it is generally not easy to control for the information conveyed by these events, and this could contribute to the relatively wide dispersion in elasticity estimates across different papers.[14]

A cleaner approach involves changes in widely tracked stock market indices. Shleifer (1986) uses changes in the S&P 500 index and the consequent demand shocks by investors tracking the index to measure the slope of the demand curve. Several other papers have followed this approach and documented a substantial price impact around S&P 500 index changes (e.g. Lynch and Mendenhall (1997)) which seems to have grown with the popularity of indexing (Morck and Yang (2001)). Similar effects have been documented for other indices in the U.S., such as the Russell indices, as well as for a variety of indices around the world. The studies for the S&P 500 suggest a price elasticity of demand of approximately unity. For example in 2000, there was an approximately 15% cumulative price impact for index additions and deletions (Petajisto (2004)) while the demand shock by mechanical indexers was approximately 10% of the shares outstanding of each stock.[15]

Clearly the actual estimates for the slope of the demand curve are not even remotely consistent with our simple CAPM calibration. It predicted only a 0.001% price impact for a 10% demand shock, and adjusting the model's parameters will not make any meaningful

[13] In fact Scholes does find a significant price effect following block trades, but it seems almost unrelated to the size of a transaction. Since the cross-sectional dispersion in the price effect is large and related to the identity of the trader, a relationship between trade size and trader identity might account for his finding. His paper does not show results within subgroups for different types of investors.

[14] A particularly amusing example of downward-sloping demand curves is provided by Rashes (2001) who finds significant price impacts even for trades where investors appeared to be confused about ticker symbols and traded a wrong stock.

[15] The size of mechanical indexers is obtained from Standard and Poor's and the Wall Street Journal, and it matches the estimates used in other papers (e.g. Blume and Edelen (2001) and Wurgler and Zhuravskaya (2002)).

changes to this enormous discrepancy. While we should not expect a perfect mapping between a simple model and reality, in this case our CAPM benchmark is obviously missing some important elements that drive the empirically observed price effect.

2.3 Suggested Alternative Hypotheses

Many different hypotheses have been suggested to explain the steep slope of demand curves for stocks in the context of index addition and deletion. Yet so far none of the papers in the literature has attempted to calibrate the commonly suggested hypotheses to actual data. Could they theoretically explain a significant fraction of the index premium? How applicable are they?

2.3.1 Liquidity

Stocks in the S&P 500 are typically among the most liquid stocks, which shows in their greater trading volume and narrower bid-ask spreads. Perhaps liquidity creates a price premium for these stocks, along the lines of Amihud and Mendelson (1986). If S&P 500 membership per se increases liquidity, this would explain at least some price impact around index changes.

However, liquidity has a much harder time explaining price effects for stocks within an index, i.e. when all stocks concerned are members of the index both before and after the event. Kaul et al. (2000) investigate an event in the Toronto Stock Exchange where the public float was officially redefined, resulting in changes in index weights across index stocks. Their estimates imply a price elasticity of demand of about -0.3.[16] Greenwood (2004) studies a large event for the Nikkei 225 index which had a significant price impact on the stocks that were in the index before and after the event. When MSCI redefined its indices (tracked closely by \$600 billion and loosely by \$3 trillion) to be based on the float and not the number of shares outstanding, many practitioners were taking speculative positions in anticipation of intra-index price effects.[17] Liquidity, as arising from index membership per se, cannot account for all these findings.

[16] This is the value of $\frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}}$ calculated by us based on the regression estimates and a 4% market share for indexers reported in the paper.

[17] "MSCI's Stock Shuffle Turns Managers Into Stock Pickers," The Wall Street Journal, 11/30/2001.

2.3.2 Market Segmentation

Merton (1987) suggests that the price of a stock is increasing in its investor base. Applying his reasoning to our context, the addition of a stock to the S&P 500 could increase its visibility to investors and make information more widely available. This could then push up the stock price.

While this explanation could contribute to the effect, it faces the same challenge as the liquidity hypothesis. It is easy to believe that the investor recognition of a stock depends on membership in the S&P 500, but it is much harder to explain why it would depend on the official weight within the index.

Instead of considering shocks to the investor base, we could also look at the increased risk aversion of active investors arising from a highly segmented market. Perhaps active investors are so poorly diversified that they cannot aggressively exploit mispricings and react to uninformed supply shocks. If we try our CAPM calibration of section 2.1 with 20 stocks instead of 1,000, we still get only a 0.05% price impact. Even this exposure to market risk is so large that it implies a very low risk aversion for investors and almost perfectly horizontal demand curves.

2.3.3 Information

Addition to the S&P 500 may convey positive information about a stock, as suggested by e.g. Denis et al. (2003). But for information to be the sole explanation, we again run into the challenge of the intra-index price effects. Other evidence can be obtained from indices such as the Russell 2000 where membership is based on a mechanical market-cap rule, and yet we still observe both economically and statistically significant price effects (e.g. Petajisto (2004)). Practitioners also keep a close eye on changes to other mechanically determined indices such as the Nasdaq 100.[18] In fact even for the S&P 500, the bureaucratic index changes in July 2002 represent a clearly uninformative event which nevertheless produced the usual magnitude for the price impact.[19]

[18] "Nasdaq 100 Index Shuffle Is Expected to Bring 13 Changes to List of Stocks," The Wall Street Journal, 11/12/2001.

[19] Barberis, Shleifer, and Wurgler (2004) point out that index membership may in fact change the beta of a stock. This could potentially lead to a price impact around index changes. However, the beta of a stock cannot change due to index membership unless mechanical fund flows are able to influence prices,

3 An Explanation with Financial Intermediaries

3.1 Motivation

Finding the fundamental value of a firm is not an easy task. It takes time and effort to investigate a firm and its environment, including the firm's products, customers, suppliers, and competitors, and this has to be done continuously as all of these may change over time. Coming up with a meaningful valuation also requires some literacy in finance. While some individual investors are certainly capable and willing to engage in this activity, it seems plausible that most of the "smart money" in the market is invested by professionals. At the end of 2000, large institutional investors accounted for 55% of the market value of stocks traded on the NYSE, AMEX, and Nasdaq, and one could argue that these institutions represent an even greater share of relatively informed investors. It may be that individual investors make the market efficient not so much by trading stocks directly but by investing part of their wealth with professional active money managers.[20]

Such institutions have emerged presumably because there is some fixed cost to becoming an informed and active market participant. End investors then pay this cost as a fee for the services provided by the professional money managers. A typical actively managed U.S. equity mutual fund charges an annual fee of approximately 1.5% of assets under management.[21] For end investors this means they should not only consider the possible mispricing of individual stocks but also whether those mispricings are large enough to justify the costs of active management.

i.e. unless demand curves slope down. Hence, any such change in beta should be taken as evidence of downward-sloping demand curves, but of course it leaves open the question about why demand curves slope down in the first place.

[20] For smaller and transitory order imbalances, it would be realistic to consider the impact of market makers on the slopes of demand curves. However, membership changes in the S&P 500 represent very large and permanent supply shocks (and their price impacts persist even after several months), so they have to be primarily accommodated by other investors with longer investment horizons. Since we are interested in price effects that last for months or years, we ignore market makers altogether.

[21] This is perhaps the most commonly quoted value for the annual fee, but there is some dispersion here. For example, Kacperczyk, Sialm, and Zheng (2004) report that the average actively managed diversified U.S. equity fund had an expense ratio of 1.28% of assets under management in 1984-1999.

3.2 The Model

We consider a setting (Figure 4) similar to the CAPM calibration in section 2.1. The main difference is an explicit layer of institutions between end investors and the stock market: the end investors can invest in the stock market only indirectly through an active manager (a stock picker) and a passive manager (who just holds the market portfolio). We also assume there are exogenous noise traders who hold a randomly chosen portfolio of stocks.[22] Since the noise traders deviate from the market portfolio, they create profitable trading opportunities for the active managers. We abstract entirely from any potential agency issues between the money managers and the end investors.

3.2.1 Assets

As before, there are N_S stocks (a large number) with a supply of 1 unit each, and a risk-free asset with an infinitely elastic supply. One period from now stock i pays a liquidating dividend of $\widetilde{x}_i = a_i + b_i \widetilde{y} + \widetilde{e}_i$ dollars. Systematic shocks to the economy are represented by the unexpected return on the market portfolio $\widetilde{y} \sim N\left(0, \sigma_m^2\right)$. Idiosyncratic shocks to the stock are denoted by $\widetilde{e}_i \sim N\left(0, \sigma_{e_i}^2\right)$. a_i and b_i are stock-specific constants. The return on the risk-free asset is normalized to zero.

To keep the mathematics simple while allowing for a large number of stocks, we make two assumptions. We let all stocks have the same values of a_i, b_i, and $\sigma_{e_i}^2$. We also assume a continuum of stocks with a measure N_S, so that our results depend on the distribution of noise trader holdings but not on their particular realizations.

3.2.2 End Investors

The economy is populated by mean-variance investors who can be aggregated into a representative investor with CARA utility and a coefficient of absolute risk aversion γ_e. Rather than investing in individual stocks, the end investor can only pick how much to invest in an actively managed portfolio and the market portfolio, with the rest of his wealth invested in

[22] Equivalently, we could assume an unobservable noisy supply for each stock. Since we will be calibrating the model to plausible parameter values, we prefer to talk explicitly in terms of noise trader holdings.

the risk-free asset. He then maximizes:

$$\begin{aligned} &\max_{\{W_a, W_{\bullet}\}} E\left[-\exp\left(-\gamma_e \widetilde{W_1}\right)\right] \\ &\text{s.t.} \quad \widetilde{W_1} = W_0 + W_a \widetilde{R}_a + W_p \widetilde{R}_m, \end{aligned} \tag{3}$$

where $\widetilde{R}_a$ and $\widetilde{R}_m$ are the excess returns on the actively managed portfolio and the market portfolio, respectively, and W_a and W_p are the dollar allocations to each.

Denoting the excess return on stock i as $\widetilde{R}_i$ and the price of the market portfolio as P_m, we can write the portfolio returns as

$$\widetilde{R}_a = \left(\sum_{i=1}^{N_S} v_i \widetilde{R}_i\right) - f_a \tag{4}$$

$$\widetilde{R}_m = \frac{1}{P_m} \sum_{i=1}^{N_S} P_i \widetilde{R}_i, \tag{5}$$

so the active portfolio has weights v_i and a constant proportional fee f_a on the portfolio return, while the market portfolio is simply a value-weighted average of individual stock returns. We can also decompose the active portfolio return into $\widetilde{R}_a = \alpha_a + \beta_a \widetilde{R}_m + \widetilde{\varepsilon}_a$ where β_a is the market beta of the portfolio and $\widetilde{\varepsilon}_a \sim N\left(0, \sigma_a^2\right)$. Then the after-fee abnormal return α_a and the idiosyncratic variance σ_a^2 of the manager's portfolio are given by:

$$\alpha_a = \sum_{i=1}^{N_S} v_i \alpha_i - f_a \tag{6}$$

$$\sigma_a^2 = \sum_{i=1}^{N_S} v_i^2 \sigma_i^2 \tag{7}$$

where α_i and σ_i^2 denote the abnormal return and the idiosyncratic variance of return for stock i.

We assume the end investor knows the expected returns and variances on the active portfolio and the passive market portfolio (but not on individual stocks). These are summary statistics of the stock market which can be learned over time in a repeated-game setting, whereas the alpha of an individual stock is randomly drawn each period and thus cannot be learned over time.

3.2.3 Active Managers

An active manager offers the end investor a portfolio with weights $\{v_i\}$ and a proportional fee f. We assume that there is a market for active managers: anyone can become an active manager by paying a fixed dollar cost C. It allows the manager to learn the stock-specific parameters a_i, b_i, and $\sigma_{e_i}^2$ and then actively pick an efficient portfolio. The manager recovers this fixed cost by imposing a fee which is a constant percentage of assets under management.[23]

Active managers compete with one another to provide the end investor with a portfolio that maximizes his expected utility (3), subject to the constraint that the managers have to earn their costs at the end investor's optimal allocation $W_a = W_a^*$.

Since we assume a fixed dollar cost but no diseconomies of scale, in equilibrium with free entry there will be only one active manager whose total fee is exactly enough to cover his fixed cost C. If the manager's fee exceeds his cost, someone else will step in, undercut the fee of the incumbent, and win the business of all end investors.[24] In reality we of course observe a large number of competing yet coexisting actively managed funds, even within relatively narrow market segments, which suggests the presence of some organizational diseconomies of scale.[25] While it would certainly be realistic to include these considerations in our model, in this paper our main objective is instead to find out how the intermediaries and their proportional fee affect the cross-sectional pricing of assets, and here a simpler structure for the mutual fund industry should help us keep our main result as clear as possible.

We allow the active managers to take both long and short positions, so the value of their stock portfolios could be zero or even negative. Yet in reality, all positions require capital – even short-only funds are constrained in their positions by the amount of capital

[23] Note that it would be very difficult to maintain any other kind of fee structure in equilibrium. Since portfolios are virtually costless to repackage, any nonlinear pricing (including nonlinear fees) would represent an arbitrage opportunity. Not surprisingly, linear fee structures also appear to be the norm in practice.

We abstract away from return-based incentive fees since 98% of U.S. mutual funds do not have such fees (Elton, Gruber, and Blake (2003)).

[24] Perhaps more realistically, we could divide the economy into n segments (industries), each with a fixed cost of $\frac{C}{n}$. In equilibrium we can then have n active managers who each specialize in one segment.

[25] Chen, Hong, Huang, and Kubik (2004) discuss the organizational diseconomies of an actively managed fund. They find empirical evidence that such diseconomies do erode fund performance. Alternative approaches are presented by Hortacsu and Syverson (2004) who suggest search costs to explain the existence of a large number of funds (including funds with different fees yet virtually identical portfolios), while Mamaysky and Spiegel (2002) suggest that multiple funds could exist to cater to investors' heterogeneous preferences.

they have. To capture this notion, we require that

$$\sum_{v_i>0} v_i = 1. \tag{8}$$

This represents a collateral requirement where all the cash generated by short sales is invested in the risk-free asset.[26] It creates the necessary link between the amount of wealth a manager controls and the size of the positions he is able to take. The dollar amount the manager invests in stock i is $W_a v_i$, so the magnitude of the v_i's is an innocuous normalization (determined together with W_a), but it is important that the manager's portfolio costs something.

We assume the management fee f is based on the combined size of the long position and the short position. The dollar fee is thus given by

$$f \sum_{i=1}^{N_S} |W_a v_i| = f W_a \sum_{i=1}^{N_S} |v_i| \,, \tag{9}$$

which translates to a fee of

$$f_a = f \sum_{i=1}^{N_S} |v_i| \tag{10}$$

as a fraction of the portfolio investment W_a. We choose this definition as an appropriate compromise for two reasons: First, it treats long and short positions symmetrically, not creating the false appearance of costless short positions. Second, if the percentage fee was only based on long positions (generating a dollar fee of fW_a), this would artificially diminish the effect of the fee in our calibrations,[27] since the manager in our model neither charges a large incentive fee (like hedge funds do) nor does he get to charge a percentage fee for a large investment in the market portfolio (like mutual funds do). In section 5.5 we explain how more realistic interpretations for the fee will magnify its effect even further.

3.2.4 Equilibrium between End Investor and Active Manager

We have to solve for the equilibrium choice variables W_a, W_p, $\{v_i\}$, and f in two stages. The end investor chooses his optimal allocations W_a and W_p, taking the excess returns on the market portfolio and the active portfolio as exogenous. We can write this problem as:

$$\max_{W_a, W_{\bullet}} E\left[u\left(W_0 + W_p \widetilde{R}_m + W_a \left(\sum_i v_i \widetilde{R}_i - f \sum_{i=1}^{N_S} |v_i| \right) \right) \right]. \tag{11}$$

[26] Investors are usually required to deposit 102% of the cash proceeds of the short sale with their broker (D'Avolio (2002)).

[27] If the dollar fee is defined simply as fW_a, the price effects in our calibrations are cut exactly in half.

Since the optimal allocations will depend on the manager's choices $\{v_i\}$ and f, we can write them as functions $W_a^*(\{v_i\}, f)$ and $W_p^*(\{v_i\}, f)$.

The active manager chooses portfolio weights $\{v_i\}$ and the fee f to maximize the same objective function, taking into account his own impact on the end investor's optimal choices $W_a^*(\{v_i\}, f)$ and $W_p^*(\{v_i\}, f)$:

$$\begin{aligned} &\max_{\{v_i\},f} E\left[u\left(W_0 + W_p\widetilde{R}_m + W_a\left(\sum_i v_i\widetilde{R}_i - f\sum_{i=1}^{N_S}|v_i|\right)\right)\right] \\ &\text{s.t.}\quad fW_a\sum_{i=1}^{N_S}|v_i| \geq C. \end{aligned} \tag{12}$$

There are two conceptual differences in the two optimization problems: First, the active manager faces the constraint that he has to cover his fixed cost C, while the end investor does not observe the constraint. Second, the end investor maximizes for exogenous values of $\{v_i\}$ and f, while the active manager realizes that the end investor's response depends on the manager's actions.[28] Hence, this is a natural case of Stackelberg equilibrium, where the manager is the Stackelberg leader who maximizes the objective function with full anticipation of the individually optimal response of the end investor.

After some algebra, we obtain the optimal allocations to the active and passive portfolios:

$$W_a^* = \frac{E\left[\widetilde{R}_a\right] - \beta_a\eta}{\gamma_e\sigma_a^2} = \frac{\alpha_a}{\gamma_e\sigma_a^2} \tag{13}$$

$$W_p^* = \frac{E\left[\widetilde{R}_m\right]}{\gamma_e\sigma_m^2} - \beta_a W_a^* = \frac{\eta}{\gamma_e\sigma_m^2} - \beta_a W_a^*, \tag{14}$$

where η denotes the market risk premium. When we plug these expressions into the active manager's maximization problem (12), we can write the objective function in terms of the certainty equivalent of the end investor:

$$W_0 + \frac{1}{2\gamma_e}\left[\underbrace{\left(\frac{\eta}{\sigma_m}\right)^2}_{\substack{\text{Sharpe ratio}\\ \text{of market}}} + \underbrace{\left(\frac{\alpha_a}{\sigma_a}\right)^2}_{\substack{\text{appraisal ratio of}\\ \text{active portfolio}}}\right]. \tag{15}$$

[28] These differences arise because the end investor represents a large number of small individual investors. Each one of them will necessarily be a price taker and will have negligible impact on the proportional fee f and the aggregate allocation W_a.

The end investor's expected utility thus depends on the Sharpe ratio of the market and the appraisal ratio of the active portfolio. In other words, problem (12) is equivalent to maximizing the appraisal ratio of the active portfolio, subject to the constraint that the manager cover the fixed cost. This portfolio advice is naturally consistent with Treynor and Black (1973) who also advocate the appraisal ratio as an appropriate objective for an active manager.

Solving the manager's problem (12), we find that the optimal portfolio weights are linear in alpha (see the appendix):

$$v_i = \left(\frac{1}{\sum_{\alpha_j > 0} \frac{\alpha_j}{\sigma_j^2}} \right) \frac{\alpha_i}{\sigma_i^2}. \tag{16}$$

Note that these are the same intuitive portfolio weights a standard CARA investor would choose.

The dollar demand of the active manager for stock i can then be expressed as

$$W_i = W_a v_i = \frac{W_a}{\sum_{\alpha_j > 0} \frac{\alpha_j}{\sigma_j^2}} \frac{\alpha_i}{\sigma_i^2} = \frac{\alpha_i}{\gamma \sigma_i^2}, \tag{17}$$

where we defined the "effective risk aversion" of the active manager as

$$\gamma = \frac{1}{W_a} \sum_{\alpha_j > 0} \frac{\alpha_j}{\sigma_j^2}. \tag{18}$$

This is the implied coefficient of absolute risk aversion of the active manager if he was a CARA investor investing his own wealth.[29] Since the manager simply invests all his assets under management in stocks, his effective risk aversion is directly determined by the end investor's dollar allocation to him. Yet this notation is very useful, as it simplifies our equations and offers a convenient interpretation in the equilibrium analysis.

3.2.5 Market Clearing

There are three groups of investors holding stock i: First, the passive manager holds the same fraction $u_p = \frac{W_{\bullet}}{P_m}$ of the supply of each stock, where P_m is the price of the market portfolio. His demand will therefore depend not on the price of stock i but on the price of the aggregate market portfolio. Second, noise traders hold a random supply $u_{in} \sim N\left(0, \sigma_u^2\right)$ which is independent of price. These are the investors who create profitable

[29] The manager's true personal risk aversion is not even defined, as he has no personal wealth or utility function.

trading opportunities for sophisticated stock pickers. Third, the active manager holds the remaining supply u_i. Thus it is the active manager whose actions will determine the cross-sectional pricing of stocks. Together, the demand of the three investors adds up to the supply of the stock:

$$u_p + u_{in} + u_i = 1. \tag{19}$$

The equilibrium price of stock i will then be

$$P_i = \underbrace{a_i}_{\substack{\text{expected} \\ \text{payoff}}} - \underbrace{b_i \eta}_{\substack{\text{discount for} \\ \text{market risk}}} - \underbrace{\gamma \sigma_{e_i}^2 u_i}_{\substack{\text{deviation} \\ \text{from CAPM}}} . \tag{20}$$

This yields an alpha which is linear in the position u_i of the active manager:

$$\alpha_i = \frac{\gamma \sigma_{e_i}^2}{P_i} u_i. \tag{21}$$

By construction, the market portfolio will always have an alpha of zero. This implies that $u_i \sim N\left(0, \sigma_u^2\right)$. In other words, the active manager will hold an equal number of shares in his long and short positions, so his exposure to market risk will automatically be zero.

We then have five remaining equilibrium variables: the allocations W_a and W_p to the active and passive managers, the market risk premium η, as well as the fee f and the effective risk aversion γ of the active manager. We also have five equations: two for the allocations, one for the portfolio value of the active manager, one for the market clearing of stock i, and one for the dollar fee. After some algebra, we obtain the following:

Proposition 1 *The equilibrium is given by:*

$$\begin{aligned}
\eta &= \frac{\gamma_e \sigma_M^2}{N_S a - \gamma_e \sigma_M^2} && (22) \\
W_p &= N_S a - \gamma_e \sigma_M^2 && (23) \\
W_a &= \frac{N_S \sigma_u}{2} \left[\sqrt{\frac{2}{\pi}} (a - b\eta) - \gamma_e \sigma_e^2 \sigma_u \right] && (24) \\
f &= \frac{C}{\sqrt{\frac{2}{\pi}} (a - b\eta) N_S \sigma_u} && (25) \\
\gamma &= \gamma_e + \sqrt{\frac{2}{\pi}} \left(\frac{a - b\eta}{\sigma_e^2 \sigma_u} \right) f. && (26)
\end{aligned}$$

Here σ_M^2 denotes the dollar variance of the market portfolio. We leave some expressions in terms of the market risk premium η to keep them simple, and we leave the last expression in terms of the endogenous variable f as we prefer to calibrate the model to a percentage fee rather than a dollar cost.

3.3 Analysis of Equilibrium

The model has essentially three free and meaningful parameters to pick: the length of the time period, the active manager's fixed cost C (which produces a fee f), and the dispersion in noise traders' holdings σ_u. For the rest of the parameters we either get reasonably good estimates from actual data (the market risk premium and volatilities) or they do not matter for our results (price normalization or the exact number of stocks). The model's restrictions then determine the joint equilibrium distributions of u_i, P_i, and α_i, which in turn determine the slope of the demand curve for a stock.

In the first calibration, we want to be as close as possible to the CAPM benchmark of section 2.1. We set the length of the period to one year, the number of stocks $N_S = 1,000$, the risk aversion of the end investors $\gamma_e = 1.247 \times 10^{-5}$ (to produce a market risk premium of $\eta = 0.05$), $a = 105$ (to normalize the average price to 100), $b = 100$ (to set the beta of the market portfolio $\beta_m = 1$), $\sigma_M^2 = 4 \times 10^8$ (to get a standard deviation of 20% for the market return), $\sigma_e^2 = 900$ (to get a standard deviation of 30% for idiosyncratic stock return), and the dispersion in noise trader holdings $\sigma_u = 0.1$ (so that the 95% confidence interval for noise trader holdings is 40% of the supply of the stock). We again investigate the price impact of an exogenous -10% supply shock which would correspond to a stock being added to the S&P 500. We then perform the same calibration with the time period set to 5 years instead of 1 year.

The expression for the effective risk aversion of the active manager perhaps most clearly reveals the unique feature of our equilibrium:

$$\gamma = \gamma_e + \sqrt{\frac{2}{\pi}} \left(\frac{a - b\eta}{\sigma_e^2 \sigma_u} \right) f. \tag{27}$$

If the fee f charged by the active manager is zero, then the active manager's risk aversion will match that of the representative end investor, and the model collapses to the CAPM benchmark. Exactly as before, a -10% supply shock to a typical stock will increase the price of the stock by only 0.11 basis points. However, the fee f has a very significant first-order effect on γ – even a tiny fee of 0.1% would increase γ by a factor of 70. Table 1 illustrates the effect of the fee on the equilibrium distribution of alphas, on the effective risk

fee	95% confidence interval for α_i	effective risk aversion γ	price impact of a -10% supply shock
0	$[-0.0022\%, 0.0023\%]$	1.25×10^{-5}	0.0011%
0.1%	$[-0.16\%, 0.16\%]$	8.99×10^{-4}	0.081%
0.5%	$[-0.79\%, 0.81\%]$	4.45×10^{-3}	0.40%
1.0%	$[-1.6\%, 1.6\%]$	8.88×10^{-3}	0.80%
1.5%	$[-2.3\%, 2.5\%]$	1.33×10^{-2}	1.20%
2.0%	$[-3.1\%, 3.3\%]$	1.77×10^{-2}	1.60%

Table 1: The effect of the management fee; one-year horizon.

annual fee	95% CI: cumulative α_i over 5 years	effective risk aversion γ	price impact of a -10% supply shock
0	$[-0.012\%, 0.012\%]$	1.25×10^{-5}	0.0062%
0.1%	$[-0.80\%, 0.82\%]$	8.15×10^{-4}	0.41%
0.5%	$[-3.8\%, 4.2\%]$	4.02×10^{-3}	2.0%
1.0%	$[-7.4\%, 8.7\%]$	8.04×10^{-3}	4.0%
1.5%	$[-11\%, 14\%]$	1.20×10^{-2}	6.0%
2.0%	$[-14\%, 19\%]$	1.61×10^{-2}	8.0%

Table 2: The effect of the management fee; five-year horizon.

aversion, and on the price impact of a -10% supply shock with a 1-year horizon. Table 2 shows the same results with a 5-year horizon.

With a 1-year horizon and a realistic fee of 1.5% of assets under management, we get a price impact of 1.20%. This is some orders of magnitude (about 1,000 times) greater than in the classical CAPM case with a zero fee. For even very small values of the fee (0.1%), the risk aversion of the end investors actually becomes *irrelevant* to the effective risk aversion of the active manager.

With a 5-year horizon, the price effects are scaled up approximately by a factor of 5. Now a -10% supply shock produces a price impact of 6%, which is economically a very significant amount and roughly equal to one half of the S&P 500 index premium. While the crucial deviation from the CAPM arises solely due to the fee, the horizon also matters a great deal if we want to get close to the empirically observed price effect.

Regardless of the horizon, our results are in stark contrast to traditional representative agent models where end investors' risk aversion shows up both in the pricing of market risk and in the pricing of idiosyncratic risk. In our setting, no such link exists. The market portfolio is still priced according to the risk aversion of the end investors, but the cross-sectional pricing of stocks is determined separately by the fee charged by the professional stock pickers.

What exactly is driving this result? The cross-sectional pricing of stocks is determined by the active manager who is constrained to invest exactly 100% of the wealth allocated to him by the end investors. In equilibrium the end investors will have to be indifferent between the actively managed portfolio, which has a positive alpha but charges a fee, and the passively managed portfolio, which has a zero alpha but also a zero fee. Hence, the before-fee alpha of the active portfolio has to be approximately equal to the fee. This in turn implies that the dispersion in the alphas of individual stocks has to be sufficiently wide in equilibrium to produce the nontrivial portfolio alpha. The dispersion in alphas thus represents an equilibrium level of "inefficiency" in the market, measured with respect to the active manager's information set.[30,31]

The alpha curve for a stock and the equilibrium distribution of alphas in the entire population of stocks are shown in figure 1. The dotted lines indicate the typical positive and negative stock positions of the manager, i.e. the ones where he just covers the fee of

[30] Here the stock market is not "efficient" in the traditional sense because an active manager can pick stocks that outperform the market. But since this outperformance cannot be obtained without a cost and in equilibrium the cost largely eliminates the gains from outperformance, we could reasonably define this market as efficient.

[31] Alternatively, we could write the stochastic discount factor of the economy as

$$\widetilde{m} = 1 - \frac{\eta}{\sigma_m^2}\widetilde{y} - \gamma \sum_{i=1}^{N_S} u_i \widetilde{e}_i.$$

The first random term accounts for the systematic discount of a stock due to market risk (the CAPM price). The remaining random terms account for the idiosyncratic mispricings of individual stocks. There is no structure to these mispricings – only the active managers conducting fundamental analysis of individual firms are able to identify them.

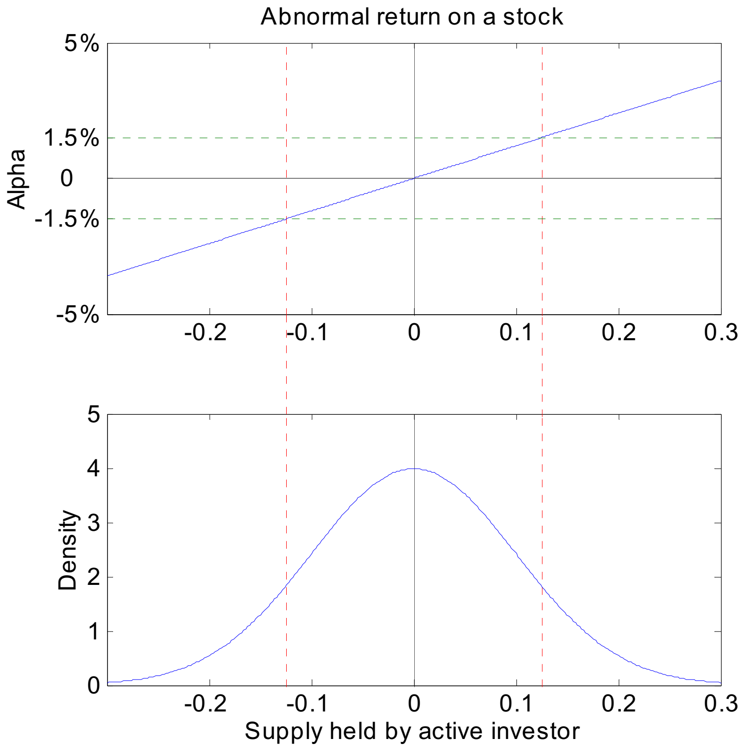

Figure 1: The alpha of an individual stock, and the distributions of alphas and active manager's stock holdings (u_i) in the entire population of stocks.

1.5%. The slope of the alpha curve is now about 1,000 times greater than in the CAPM benchmark.

Regardless of the horizon, the distribution of annual alphas is the same. Yet the pricing results are very different, because the alphas across the entire period are capitalized into prices today (figure 2). With a 1-year horizon, a 1% annual alpha translates to a 1% underpricing, but with a 5-year horizon the same 1% annual alpha translates to a 5% underpricing.

How should we interpret the horizon of the model? In a one-period model, the horizon is essentially a period of time after which prices fully converge to their fundamental values. Yet in reality, no such convergence is guaranteed for stocks. It is thus better to think in terms of the expected half-life of a mispricing – e.g., the 5-year horizon should be interpreted as an expected half-life of 2.5 years for a mispricing.

The choice of an appropriate horizon then becomes primarily an empirical question.

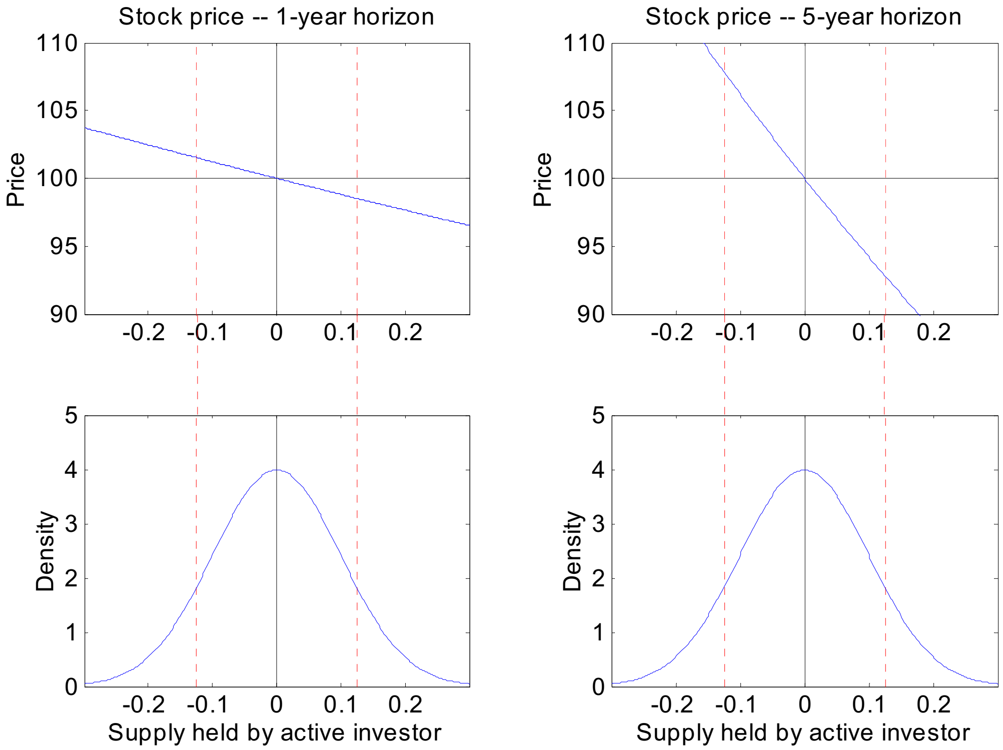

Figure 2: The price of an individual stock, and the distributions of prices and active manager's stock holdings in the entire population of stocks, for 1-year and 5-year horizons. The distribution of annual alphas is the same in each case, but since the alphas over the entire period are capitalized into prices today, the longer horizon scales up the "mispricings" today.

DeBondt and Thaler (1985) find slow mean reversion in returns over a 3-5-year period, while Jegadeesh and Titman (2001) find momentum at a 1-year horizon and a partial or full reversal (depending on the sample period) over the following 4 years. This suggests that mispricings may indeed take several years to reverse. Cohen, Gompers, and Vuolteenaho (2002) construct a VAR model which allows them to estimate the reversal of a pure expected-return shock. Their results indicate a half-life of at least 2.5 years (figures 5 and 6 in the appendix).

In the context of index changes, the price impact seems to last at least for 2 months, but beyond that our tests start to lose power to distinguish between alternatives. Professional investors seem to have divergent views on this topic, with some of them believing a stock will have a permanent premium as long as it stays in the index. Certainly a full 1-year reversal seems implausible, as it would offer easy opportunities to earn 15% annual alphas.[32] More generally, if a moderate mispricing can exist today, how can we be so sure it cannot exist tomorrow?

Overall, a half-life of 2.5 years for a mispricing seems roughly consistent with empirical evidence, so we adopt the 5-year horizon as a reasonable compromise. The main virtue of the 1-year horizon is that it makes the numbers in the calibration a little more transparent.

The model's predicted 6% price impact for S&P 500 index changes should still be considered a lower bound for the effect we describe. We have assumed frictionless short selling, and consequently the actively managed portfolio turned out to be a market-neutral long-short portfolio. In reality mutual funds almost never take short positions, and they carry heavy exposure to systematic market risk. In section 5.5 we discuss how to interpret our model in a more realistic context. It turns out this further increases the predicted price effect – with plausible parameter values we get a 15% effect for S&P 500 index changes. While we should not expect our simple model to be an accurate predictor of real-life price impact, the numbers from our calibrations should be taken as evidence that the mechanism we describe is economically significant and has the potential to explain a large part of the empirically observed price effects.

4 Empirical Implications

4.1 Predictions

The most immediate testable prediction of the model is the overall magnitude of the slopes of demand curves under reasonable parameter values. This was already discussed in the numerical calibration of the previous section.

Most of the model's other testable implications stem from two equations:

$$\begin{aligned} P_i &= a_i - b_i\eta - \gamma\sigma_{e_i}^2 u_i && (28) \\ \gamma &= \gamma_e + \sqrt{\frac{2}{\pi}}\left(\frac{a-b\eta}{\sigma_e^2\sigma_u}\right)f. && (29) \end{aligned}$$

The price of a stock is given by its CAPM price $(a_i - b_i\eta)$ minus a deviation $(\gamma\sigma_{e_i}^2 u_i)$ due

[32] There is anecdotal evidence that hedge funds tend to bet in favor of the index effect, taking advantage of its drift between the announcement and effective days, as opposed to betting against it.

to idiosyncratic risk.[33] As the equilibrium holdings (u_i) of the active manager change, the price impact is given by the dollar variance ($\sigma^2_{e_i}$) of the stock's payoff times the effective risk aversion (γ) of the active manager. The price elasticity of demand for stock i is then

$$\frac{\frac{dQ_i}{Q_i}}{\frac{dP_i}{P_i}} = \frac{\frac{du_i}{1}}{\frac{dP_i}{P_i}} = P_i \frac{du_i}{dP_i} = -\frac{P_i}{\gamma \sigma^2_{e_i}}. \tag{30}$$

Implication 1 *The demand curve is steeper for stocks with greater idiosyncratic risk.*

The effective risk aversion of the active manager is supposed to be the same across all stocks. However, if the stock market is segmented so that each active manager (stock picker) generally focuses on a subset of the available stocks,[34] we may also see some variation in the manager's effective risk aversion as his fee changes from one segment to another.

Implication 2 *The demand curve is steeper for stocks in segments of the market with a greater fee for active management.*

Implication 3 *The demand curve is steeper for stocks in segments of the market with a greater cost of information acquisition.*

The latter implication holds when the fee for active management is related to the information acquisition cost of the manager.

Implication 4 *The demand curve is steeper for stocks in segments of the market with less dispersion in noise trader holdings.*

It may be somewhat surprising that a larger dispersion of noise trader holdings actually makes demand curves more horizontal and in that sense makes the market more efficient. The reason is that the equilibrium dispersion of alphas across stocks has to be the same as the active managers still earn their fees, but now the same dispersion of alphas exists over a wider range of the managers' stock holdings, so the change in alpha (and price) for a supply

[33] Note that the deviation is sometimes positive and sometimes negative (depending on the sign of u_i), so idiosyncratic risk alone will not be linked to expected returns.

[34] In fact, if there is no segmentation, then small firms (measured by operating size such as revenues) will always command a smaller risk premium in equilibrium, giving rise to an inverse size effect. When the market is segmented, it is possible to maintain a relatively constant density of investors in each stock. We address these issues explicitly in a separate appendix to this paper.

shock of a given size is smaller. In other words, a noise trader can minimize his own price impact by trading in stocks where the volatility of aggregate noise trader holdings is high.

Our model also implies that noise traders can move prices, and in fact they can increase the volatility of a stock beyond the volatility of its fundamentals.

Implication 5 *Stocks with a greater volatility of noise trader holdings will exhibit greater price volatility, unless the shocks to noise trader holdings are inversely correlated with fundamental news.*

4.2 Evidence

The link between active management fees and the slopes of demand curves is tested in a separate paper (Petajisto (2004)), which provides empirical evidence from the large-cap and small-cap segments of the market using data from S&P 500 and Russell 2000 index changes. It finds that small-cap stocks exhibit steeper demand curves than large-cap stocks, which is consistent with the higher management fees of active small-cap mutual funds. Naturally, it would be interesting to test this prediction even more broadly across various market segments or multiple countries.

The predicted cross-sectional link between idiosyncratic risk and demand curves is strongly confirmed by empirical tests for both indices (Petajisto (2004)).

5 Interpretations and Further Discussion

5.1 Welfare Analysis: Institutions Maximize Efficiency of Prices

5.1.1 Measures of Welfare

Let us turn to a broader question: Given an exogenous fixed cost C to actively managing a stock portfolio, what would be the optimal market design for investors? Does it resemble the one we assumed in our model?

The appropriate group for welfare analysis are the noise traders.[35] How much are they hurt because of their randomly chosen portfolio? This is determined directly by the slopes

[35] Since there is an exogenous positive supply of risky assets, the rational investors could potentially collude by withholding their investments and thus driving expected dollar returns to infinity. Thus their expected utility expressed in terms of certainty equivalent is unbounded, and welfare analysis for them is not meaningful.

of the demand curves for stocks – the steeper the demand curves, the more the noise trades will push prices against themselves. For example, anyone who needs to trade a stock for liquidity reasons would like to be facing as flat demand curves as possible. Moreover, the flatter the demand curves, the closer prices will be to their fundamental values. The slopes of demand curves will therefore also be a good measure of the informational efficiency of market prices.

5.1.2 Maximizing Welfare across All Market Designs

The optimization problem is then to minimize the slopes of demand curves, subject to the constraint that rational investors are still willing to hold market-clearing quantities of stocks in the cross-section while bearing the cost C.

The payoff of stock i is given by $\widetilde{x}_i = a + b\widetilde{y} + \widetilde{e}_i$, and its supply available to active investors is again u_i. The payoff of the entire active portfolio when we include the fixed cost C is given by

$$\widetilde{X} = \left[\sum_{i=1}^{N_S} u_i \widetilde{x}_i\right] - C = \left[\sum_{i=1}^{N_S} u_i \left(a + b\widetilde{y} + \widetilde{e}_i\right)\right] - C. \tag{31}$$

The price of the portfolio is $P = \sum_{i=1}^{N_S} u_i P_i$.

Since the noise traders take the opposite side of the trade against active investors, they want to maximize the price of this portfolio. Hence, the welfare maximization problem is:

$$\begin{aligned} \max_{P,\gamma_c} \quad & P \\ \text{s.t.} \quad \frac{E\left[\widetilde{X}\right] - P}{\gamma_c Var\left[\widetilde{X}\right]} & \geq 1 \\ \gamma_c & \geq \gamma_e. \end{aligned} \tag{32}$$

The collective absolute risk aversion of all investors is γ_e. A subset of them with a collective absolute risk aversion γ_c chooses to bear the risk of the active portfolio. The first constraint requires that the demand by this subset of investors for the active portfolio has to be at least equal to its supply of 1 unit. The second constraint points out that since this group is a subset of all investors, its absolute risk aversion cannot be lower than that of all investors.

At the optimum, both constraints clearly have to be binding. Hence, we find that

$$E\left[\widetilde{X}\right] - \gamma_e Var\left[\widetilde{X}\right] = P. \tag{33}$$

Given the same noise trader holdings as before, we have $u_i \sim N\left(0, \sigma_u^2\right)$. The systematic parts of the payoffs cancel out and we can write:

$$\begin{aligned} \widetilde{X} &= \left[\sum_{i=1}^{N_S} u_i \widetilde{e}_i\right] - C && (34) \\ E\left[\widetilde{X}\right] &= -C && (35) \\ Var\left[\widetilde{X}\right] &= \sum_{i=1}^{N_S} u_i^2 \sigma_e^2 = N_S \sigma_u^2 \sigma_e^2. && (36) \end{aligned}$$

Hence, the optimum price of the portfolio will have to satisfy:

$$P = -C - \gamma_e N_S \sigma_u^2 \sigma_e^2. \tag{37}$$

To check the optimality of our model with institutions, we compute the price of the active portfolio using the equilibrium stock prices (20) we found earlier:

$$P = \sum_{i=1}^{N_S} u_i P_i = \sum_{i=1}^{N_S} u_i \left(a - b\eta - \gamma \sigma_{e_i}^2 u_i\right) = -\gamma \sum_{i=1}^{N_S} u_i^2 \sigma_{e_i}^2 = -\gamma N_S \sigma_u^2 \sigma_e^2. \tag{38}$$

We find the value of γ from equations (25) and (26):

$$\gamma = \gamma_e + \frac{C}{N_S \sigma_u^2 \sigma_e^2}. \tag{39}$$

Thus the price in our model will be

$$P = -\left(\gamma_e + \frac{C}{N_S \sigma_u^2 \sigma_e^2}\right) N_S \sigma_u^2 \sigma_e^2 = -C - \gamma_e N_S \sigma_u^2 \sigma_e^2, \tag{40}$$

which turns out to be equal to the welfare-maximizing price (37).

Hence, the optimal market design which produces the flattest demand curves for a fixed cost C is precisely the structure we assumed in the model. Our institutional setting is the way to implement this market design.[36]

5.1.3 Endogenous Institutions

There is a simple reason why the presence of financial institutions is an optimal market design. The institutions allow every investor in the economy to share the risk of the active portfolio, which in turn allows more aggressive trading against the noise traders. The proportional fee maintains this equilibrium because it prevents the investors from trading

[36] See the appendix for a further discussion of optimal market designs.

so aggressively that alphas would be driven to zero. The equilibrium alphas are just large enough to cover the fixed cost C, and due to efficient risk-sharing there are no additional costs.[37]

Hence, far from being a friction in our model, institutions actually help us mitigate the friction arising from the underlying cost. They bring prices closer to the neoclassical CAPM benchmark, thus making the market more informationally efficient.[38]

5.2 Grossman and Stiglitz (1980) and the Necessity of Institutions

Our basic economic story with an "equilibrium degree of disequilibrium" is very much in the spirit of the insightful paper by Grossman and Stiglitz (1980).[39] Could we perhaps use their model to explain downward-sloping demand curves?

Grossman and Stiglitz present a single-asset model with informed investors, uninformed investors, and noise traders. The informed traders observe a signal of the fundamental value of the asset. The uninformed investors use the price of the asset to infer the signal of the informed, but the inference is noisy due to the unobserved holdings of noise traders. An uninformed investor can also become informed by paying a certain cost. The fraction of investors who choose to become informed is determined endogenously, so that in equilibrium

[37] The certainty equivalent of the CARA investors for investing in the active portfolio is $\frac{1}{2}\gamma_e N_S \sigma_u^2 \sigma_e^2$. This is independent of the cost, and since $C = 0$ represents the CAPM, this means that the rational investors are indifferent between our model and a standard CAPM world. The cost does not matter to them because in our model demand curves will be steeper to let the investors exactly cover the cost.

[38] This discussion is about demand curves in general. In the context of the S&P 500, one might ask why there is no actively managed fund that exclusively takes positions against index funds around index changes, especially since these events are almost costless to identify.

We can only speculate on possible answers: If such index changes were very common, we would indeed expect such funds to emerge, but the current number of index changes may still be too small to justify establishing a separate fund for that purpose only. Since the price effects may last for many years (the empirical evidence here has been inconclusive), the annual alpha for such a fund need not be so dramatically high. The fund would also not be very well diversified – in particular it would be exposed to the systematic risk of the index premium going up, which is exactly what happened in the 1990s as large amounts of money flowed into S&P 500 index funds.

[39] In fact, Allen and Gale (1994) come remarkably close to Grossman and Stiglitz (1980). Similarly, the most fundamental difference between our paper and theirs (namely, the delegation of portfolio management) is the same.

the investors are indifferent between the two choices. The cost of becoming informed determines the equilibrium level of "inefficiency" in the market.

Part of the reason demand curves slope down in that model is that the uninformed investors cannot distinguish whether a supply shock came from the informed traders (because they received good news about the stock) or the noise traders (conveying no information about the stock). However, we are concerned about demand curves for stocks in the absence of new information. For example, when a stock is added to the S&P 500, every active trader in the stock who is not consciously ignoring news will know who the new buyers are and why the stock price went up. Thus any price effect from index addition would have to come from the risk aversion of the investors and not the rational expectations story of the model.

Let us then investigate a multi-asset generalization of the Grossman-Stiglitz model to see if it would fit better. Assume that in a large cross-section of stocks, the uninformed investors are completely passive and thus have a perfectly inelastic demand.[40] Prices are then exclusively set by the informed investors.

To generate the same slope for the demand curve as in our model with a fee of 1.5%, the informed investors would have to have a collective risk aversion equal to the effective risk aversion of our active manager (Table 2 on page 21). Since this is about 1,000 times the absolute risk aversion of all investors in the economy, it implies that one investor out of 1,000 would choose to become informed. Essentially this investor faces a trade-off: either he is uninformed and holds a tiny fraction of the market portfolio, or he becomes informed and suddenly takes large enough positions to accommodate all the demand shocks due to noise traders.

It seems like a stretch to say that this huge increase in his risky portfolio (almost 100-fold in our calibration) comes from the investor's personal wealth or personal borrowing which would require collateral. Instead we could interpret this more plausibly as the investor becoming an informed intermediary who primarily invests other people's money.

[40] When the cross-section of firms exhibits wide dispersion in operating sizes and scaled-price ratios, these simple measures become virtually useless for the time-series trading of an individual stock. Without more detailed stock-specific information, the uninformed investors can therefore only have an almost perfectly inelastic demand for an individual stock.

Certainly the incentive of an informed investor to sell money-management services to others is considerable.

This takes us to the central issue: once the investor starts investing other people's money, we can no longer use his personal risk aversion to explain his investment behavior! His effective risk aversion would now be determined by how much wealth other investors are willing to allocate to him. Yet the Grossman-Stiglitz setting effectively assumes even the informed investors still keep investing their own wealth but they just borrow massively to finance their very large portfolios. Thus the model is missing the crucial part of the mechanism which is the trade-off of end investors (uninformed investors) when allocating wealth to active managers (informed investors) and the resulting equilibrium value for the effective risk aversion of the active managers.

Hence, to answer our question about equilibrium slopes of demand curves, we do indeed need something like our model where the delegation of portfolio management is made explicit. Costly information acquisition, conducted by individual investors directly, would be very hard to reconcile with a plausible multi-asset equilibrium.

5.3 Transaction Costs

Could we perhaps interpret the management fee in our model as a transaction cost that the representative investor has to pay when trading individual stocks? Would this produce results similar to our setup with financial intermediaries?

The first immediate challenge for transaction costs is their magnitude. Stocks added to the S&P 500 typically have a market capitalization of several billion dollars. Transaction costs for turning around a position in such mid-cap and large-cap stocks can even be less than 0.1%. Yet the S&P 500 premium is about 15%, which is certainly sufficient to produce abnormal returns even net of transaction costs. Moreover, some of the largest additions such as Goldman Sachs, UPS, and Microsoft have had the lowest transaction costs, yet they have experienced some of the largest price impacts.

The more fundamental challenge is that when end investors trade stocks directly, they will very aggressively exploit any alphas net of transaction costs, again due to the low risk aversion implied by the market risk premium, so that in equilibrium such abnormal returns cannot exist. Yet empirical evidence on demand curves shows that prices (and alphas) change smoothly even beyond the transaction cost as we vary the size of the supply shock. A story based on transaction costs cannot match this key feature of demand curves exhibited by our model.[41]

5.4 CRRA Utility

In a multi-asset setting, the normal distribution for stock returns combined with CARA utility offers by far the greatest analytical convenience. However, this may leave us wondering whether our results are robust to other forms of utility functions such as CRRA utility.

The conceptual difference introduced by CRRA utility is that when an investor evaluates an additional idiosyncratic gamble, his local coefficient of absolute risk aversion will now be random. It depends on his future level of wealth which is determined by the random outcomes of his other investments. The outcome of systematic market shocks overwhelmingly dominates idiosyncratic random shocks – in our model with 1,000 stocks the dollar variance of aggregate market risk is about 50,000 times the dollar variance of the aggregate actively managed portfolio. Hence, we can think of the market shock as determining the representative investor's level of future wealth and thus his future local coefficient of absolute risk aversion, and then evaluate idiosyncratic gambles assuming a constant (yet randomly drawn) level of local risk aversion.

We investigate two separate cases: the CAPM calibration (section 2.1) and our model itself (section 3). We assume the investor's entire wealth is invested in the market portfolio, which provides an upper bound for any wealth effects.

We start by assuming the market return is lognormally distributed. This makes the investor's future wealth and hence the future local coefficient of risk aversion also lognormally distributed. We then approximate the coefficient of absolute risk aversion by a normal distribution. Using this random risk aversion, we derive the investor's demand for a small idiosyncratic payoff. If this approximated equilibrium demand by a CRRA investor is the

[41] Introducing heterogeneity into the beliefs of investors would not make transaction costs a more plausible explanation. In equilibrium the end investors would be able to disagree about the value of a stock only within the narrow bands of the transaction cost; otherwise they would take extreme positive and negative positions in individual stocks (far beyond anything we observe in the real market).

Similarly, any attempt to obtain large price effects from investor disagreement in the model of Fama and French (2004) will face the same issue of counterfactually large short interest in individual stocks.

same as that by a regular CARA investor, then both the CARA and CRRA utility must also produce the same equilibrium prices.

In the CAPM calibration, a numerical calculation with a one-year horizon verifies that the demand for an idiosyncratic payoff by a CRRA investor is virtually identical to that by a CARA investor – the negligible price impact of 0.11 bp for a 1-year horizon increases by only 4% (0.004 bp), while the 5-year price impact increases by 20%, both of which are meaningless in absolute terms (see the appendix). In the model itself, we simply replace the idiosyncratic payoff of a single stock with the idiosyncratic payoff of the actively managed portfolio, and we again get the same result for both utility specifications. This is perhaps not surprising, because the wealth effects induced by CRRA utility can only show up when there is very large variation in the investor's future wealth.[42]

5.5 Interpretation of Long-Short Positions

How should we think about our model in a realistic world where only a small fraction of investors ever take short positions?

In our earlier calibration, the 95% confidence interval for the holdings of both the noise traders and the active managers was $[-20\%, 20\%]$ of the supply of the stock, while the passive manager held exactly 100% of the market portfolio. If we simply shift 20% of the market portfolio to the active managers and 20% to the noise traders, the individual stock positions begin to look more reasonable, as the noise traders and the active managers would short only about 2% of the stocks. The active managers would hold large positions in the market portfolio, but they would also be benchmarked against it and their alpha would still be derived from the long-short portfolio on top of the market portfolio. Moreover, if the managers still charge the same fee of assets under management, in this case the managers' effective fee for the same long-short portfolio would be increased by a factor of about 2.5 (relative to the earlier calibration). This would scale up the slopes of the demand curves by the same factor, so the effects of the intermediaries would become more prominent – in

[42] There is also a perhaps simpler intuition for the irrelevance of the utility specification. In a continuous-time setting, the dollar demand by a CRRA investor for an idiosyncratic asset is given by $\frac{W}{\gamma_R} \times \frac{\mu}{\sigma^2}$, where W is the investor's wealth, γ_R is his coefficient of relative risk aversion, and μ and σ are the instantaneous drift and volatility of the asset. This looks exactly like the dollar demand of a CARA investor in a discrete-time setting, which is given by $\frac{1}{\gamma} \times \frac{\mu}{\sigma^2}$. Since both investors have the same demand function for idiosyncratic payoffs, also the slopes of these demand curves must be the same.

fact it would turn the predicted 6% price impact to a 15% price impact, which is almost identical to the current S&P index premium.

While such a close match between the theoretically predicted effect and the empirically observed effect is likely to be a coincidence, it is important that the actual effect is approximately within range of plausible parameter values for the model. Since that is certainly the case with our model, it means that our explanation has the potential to be the dominant economic force behind the effect.

5.6 Other Types of Firms

The critical feature in our story is that the investors bearing market risk cannot also be the ones doing the cross-sectional pricing of stocks, because those two activities imply very different levels of risk aversion. In this context, how should one think about firms such as investment banks with large investment portfolios of their own? They should be sophisticated institutions which are capable of active trading in individual stocks, yet they still sometimes bear significant exposure to market risk.[43]

An investment bank with a proprietary trading portfolio can essentially be considered a closed-end fund. It actively trades individual stocks and the trading profits are equally distributed among shareholders. The costs of such a trading operation are reflected in the expenses of the firm and they are also equally distributed among shareholders, acting like a percentage fee on assets under management. In a competitive equilibrium we would expect the firm to raise capital by issuing shares until the abnormal return on the capital is approximately equal to the firm's costs. Hence, it makes no difference for our model whether the active managers run open-end mutual funds, closed-end mutual funds, or public corporations with proprietary trading portfolios.

However, it remains a puzzle why such an investment firm would simultaneously choose a very large exposure to market risk and very small exposure to idiosyncratic risk. This apparently schizophrenic attitude toward risk could result from benchmarking – just like an actively managed mutual fund, the investment firm can ignore market risk and let the end investors choose their own exposure to it. In the presence of short-sales costs, it may

[43] Large university endowments may also be considered potential candidates for this category. However, such endowments tend to be invested through intermediaries who collect fees, which makes the endowments themselves look like end investors rather than active managers with capital.

indeed be optimal for active managers to combine their long-short equity portfolios with large positions in the market portfolio.

6 Conclusions

In a standard neoclassical multi-asset setting such as the CAPM, both the market risk premium and the slope of the demand curve for an individual stock are jointly determined by the risk aversion of the representative investor. If we back out the representative investor's risk aversion from any empirically plausible market risk premium, we find a relatively low implied risk aversion; if we back it out from the empirically observed slope of the demand curve for an individual stock, we find a relatively high implied risk aversion. The two estimates differ by several orders of magnitude, presenting us with a well-known puzzle in finance.

In this paper we propose an explanation for the puzzle. In traditional representative agent models it is implicitly assumed that financial intermediaries have no meaningful effect on prices so that we can ignore them and let the owners of wealth invest directly in the stock market. However, this may not be an innocuous assumption. When most of the informed active investors are professional money managers who do not own the wealth they invest, the slope of the demand curve for a stock is determined by how much wealth they are given to manage. Since the active managers charge a fee for their services, the amount of wealth they manage and hence the slopes of demand curves are determined almost entirely by the fee and not by anyone's risk aversion.

This result arises from a straightforward intuition: in equilibrium, the active managers have to approximately earn their fees. Thus there persists an equilibrium level of market "inefficiency" which allows the active managers to recover what are presumably their fixed costs for acquiring information and actively trading on it. This severs the link between risk aversion and the demand curves for individual stocks. In contrast, the risk premium on the aggregate market portfolio is still entirely set by the end investors' risk aversion since the broad asset allocation decision between stocks and bonds is a decision they make directly.

The magnitude of this effect can be surprisingly large. In our calibration, increasing the annual fee from zero, which corresponds to the CAPM benchmark, to 1.5% can increase the slope of the demand curve by a factor of 1,000. With a five-year horizon, this fee may increase the price impact of the S&P 500 index membership shock from less than one basis point to a very significant 6.0%. When we allow active managers to hold market risk and be benchmarked against it, as in the real mutual fund industry, the price impact increases to 15%.

We believe this paper makes two main contributions. It suggests a generally applicable explanation to the persistent puzzle about downward-sloping demand curves, producing not only the correct sign for the effect but also the correct order of magnitude. More broadly, it provides a concrete illustration that the presence of financial institutions does have pricing implications, even without agency issues, reinforcing the conclusions of Ross (1989) and Allen (2001) about the relevance of institutions in asset pricing.

References

[1] Allen, F., 2001, "Do Financial Institutions Matter?" *Journal of Finance*, vol. 56, no. 4, 1165-1175.

[2] Allen, F. and D. Gale, 1994, "Limited Market Participation and Volatility of Asset Prices," *American Economic Review*, vol. 84, no. 4, 933-955.

[3] Barberis, N. and A. Shleifer, 2002, "Style Investing," *Journal of Financial Economics*, vol. 68, no. 2, 161–199.

[4] Barberis, N., A. Shleifer, and J. Wurgler, 2004, "Comovement," *Journal of Financial Economics*, forthcoming.

[5] Barberis, N. and R. Thaler, 2003, "A Survey of Behavioral Finance," *Handbook of the Economics of Finance.*

[6] Berk, J.B., 1995, "A Critique of Size-Related Anomalies," *Review of Financial Studies*, vol. 8, no. 2, 275-286.

[7] Berk, J.B. and R.C. Green, 2002, "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, vol. 112, no. 6, 1269-1295.

[8] Black, F., 1986, "Noise," *Journal of Finance*, vol. 41, no. 3, 529-543.

[9] Blume, M.E. and R.M. Edelen, 2001, "On S&P 500 Index Replication Strategies," working paper, University of Pennsylvania.

[10] Chen, H., G. Noronha, and V. Singal, 2004, "The Price Response to S&P 500 Index Additions and Deletions: Evidence of Asymmetry and a New Explanation," *Journal of Finance*, vol. 59, vol. 4, 1901–1929.

[11] Chen, J., H. Hong, M. Huang, and J.D. Kubik, 2004, "Does Fund Size Erode Performance? Organizational Diseconomies and Active Money Management," *American Economic Review*, forthcoming.

[12] Chen, J., H. Hong, and J.C. Stein, 2002, "Breadth of Ownership and Stock Returns," *Journal of Financial Economics*, vol. 66, no. 2-3, 171-205.

[13] Chordia, T., 2001, "Liquidity and Returns: The Impact of Inclusion into the S&P 500 Index," working paper.

[14] Cohen, R., P. Gompers, and T. Vuolteenaho, 2002, "Who Underreacts to Cash-Flow News? Evidence from Trading between Individuals and Institutions," *Journal of Financial Economics*, vol. 66, no. 2-3, 409-462.

[15] Daniel, K., M. Grinblatt, S. Titman, and R. Wermers, 1997, "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks," *Journal of Finance*, vol. 52, no. 3, 1035-1058.

[16] Daniel, K., D. Hirshleifer, and A. Subrahmanyam, 2001, "Overconfidence, Arbitrage, and Equilibrium Asset Pricing," *Journal of Finance*, vol. 56, no. 3, 921-965.

[17] D'Avolio, G., 2002, "The Market for Borrowing Stock," *Journal of Financial Economics*, vol. 66, no. 2-3, 271-306

[18] De Bondt, W.F.M. and R. Thaler, 1985, "Does the Stock Market Overreact?" *Journal of Finance*, vol. 40, 793-805.

[19] De Long, J.B., A. Shleifer, L.H. Summers, and R.J. Waldmann, 1990, "Noise Trader Risk in Financial Markets," *Journal of Political Economy*, vol. 98, no. 4, 703-738.

[20] Denis, D.K., J.J. McConnell, A.V. Ovtchinnikov, and Y. Yu, 2003, "S&P 500 Index Additions and Earnings Expectations," *Journal of Finance*, vol. 58, no. 5, 1821-1840.

[21] Edmister, R., A.S. Graham, and W.L. Pirie, 1994, "Excess Returns of Index Replacement Stocks: Evidence of Liquidity and Substitutability," *Journal of Financial Research*, vol. 17, no. 3, 333-346.

[22] Elton, E.J., M.J. Gruber, and C.R. Blake, 2003, "Incentive Fees and Mutual Funds," *Journal of Finance*, vol. 58, no. 2, 779-804.

[23] Fama, E.F. and K.R. French, 2004, "Disagreement, Tastes, and Asset Prices," working paper, University of Chicago.

[24] Glosten L.R. and P.R. Milgrom, 1985, "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, vol. 14, no. 1, 71-100.

[25] Gomes, J., L. Kogan, and L. Zhang, 2003, "Equilibrium Cross Section of Returns," *Journal of Political Economy*, vol. 111, no. 4, 693-732.

[26] Gompers, P.A. and A. Metrick, 2001, "Institutional Investors and Equity Prices," *Quarterly Journal of Economics*, 229-259.

[27] Greenwood, R., 2004, "Short- and Long-Term Demand Curves for Stocks: Theory and Evidence on the Dynamics of Arbitrage," *Journal of Financial Economics*, forthcoming.

[28] Grossman, S. and J.E. Stiglitz, 1980, "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, vol. 70, 393-408.

[29] Grossman, S. and M. Miller, 1988, "Liquidity and Market Structure," *Journal of Finance*, vol. 43, no. 3, 617-637.

[30] Gruber, M.J., 1996, "Another Puzzle: The Growth in Actively Managed Mutual Funds," *Journal of Finance*, vol. 51, no. 3, 783-810.

[31] Holthausen, R.W., R.W. Leftwich, and D. Mayers, 1987, "The Effect of Large Block Transactions on Security Prices: A Cross-Sectional Analysis," *Journal of Financial Economics*, vol. 19, no. 2, 237-267.

[32] Holthausen, R.W., R.W. Leftwich, and D. Mayers, 1990, "Large-Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects," *Journal of Financial Economics*, vol. 26, no. 1, 71-95.

[33] Hortacsu, A. and C. Syverson, 2004, "Search Costs, Product Differentiation, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," *Quarterly Journal of Economics*, vol. 119, no. 2, 403-456.

[34] Jegadeesh, N. and S. Titman, 2001, "Profitability of Momentum Strategies: An Evaluation of Alternative Explanations," *Journal of Finance*, vol. 56, no. 2, 699-720.

[35] Kacperczyk, M.T., C. Sialm, and L. Zheng, 2004, "On Industry Concentration of Actively Managed Funds," *Journal of Finance*, forthcoming.

[36] Kaul, A., V. Mehrotra, and R. Morck, 2000, "Demand Curves for Stocks *Do* Slope Down: New Evidence from an Index Weights Adjustment," *Journal of Finance*, vol. 55, no. 2, 893-912.

[37] Kyle, A.S., 1985, "Continuous Auctions and Insider Trading," *Econometrica*, vol. 53, no. 6, 1315-1335.

[38] Loderer, C., J.W. Cooney, and L.D. van Drunen, 1991, "The Price Elasticity of Demand for Common Stock," *Journal of Finance*, vol. 46, no. 2, 621-651.

[39] Lynch, A. and R. Mendenhall, 1997, "New Evidence on Stock Price Effects Associated with Changes in the S&P 500 Index," *Journal of Business*, vol. 70, no. 3, 351-383.

[40] Mamaysky, H. and M. Spiegel, 2001, "A Theory of Mutual Funds: Optimal Fund Objectives and Industry Organization," working paper, Yale School of Management.

[41] Merton, R.C., 1987, "A Simple Model of Capital Market Equilibrium with Incomplete Information," *Journal of Finance*, vol. 42, no. 3, 483-510.

[42] Merton, R.C. and Z. Bodie, 2002, "The Design of Financial Systems: Towards a Synthesis of Function and Structure," working paper.

[43] Mitchell, M., T. Pulvino, and E. Stafford, 2004, "Price Pressure around Mergers," *Journal of Finance*, vol. 59, no. 1, 31-63.

[44] Morck, R. and F. Yang, 2001, "The Mysterious Growing Value of S&P 500 Membership," NBER working paper.

[45] Nanda, V., M.P. Narayanan, V.A. Warther, 2000, "Liquidity, Investment Ability, and Mutual Fund Structure," *Journal of Financial Economics*, vol. 57, 417-443.

[46] Petajisto, A., 2004, "The Index Premium and Its Implications for Index Funds," Yale ICF working paper.

[47] Rashes, M.S., 2001, "Massively Confused Investors Making Conspicuously Ignorant Choices (MCI–MCIC)," *Journal of Finance*, vol. 56, no. 5, 1911–1927.

[48] Ross, S., 1976, "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, vol. 13, 341-360.

[49] Ross, S., 1989, "Institutional Markets, Financial Marketing, and Financial Innovation," *Journal of Finance*, vol. 44, no. 3, 541-556.

[50] Scholes, M., 1972, "The Market for Securities: Substitution versus Price Pressure and the Effects of Information on Share Price," *Journal of Business*, vol. 45, 179-211.

[51] Sharpe, W.F., 1981, "Decentralized Investment Management," *Journal of Finance*, vol. 36, no. 2, 217-234.

[52] Shleifer, A., 1986, "Do Demand Curves for Stocks Slope Down?" *Journal of Finance*, vol. 41, no. 3, 579-590.

[53] Shleifer, A. and R.W. Vishny, 1997, "The Limits of Arbitrage," *Journal of Finance*, vol. 52, no. 1, 35-55.

[54] Treynor, J.L. and F. Black, 1973, "How to Use Security Analysis to Improve Portfolio Selection," *Journal of Business*, vol. 46, no. 1, 66-86.

[55] Wurgler, J. and K. Zhuravskaya, 2002, "Does Arbitrage Flatten Demand Curves for Stocks?" *Journal of Business*, vol. 75, no. 4, 583-608.

A Derivation of Formulas

A.1 CAPM Benchmark

We write the representative CARA investor's problem in the mean-variance form as

$$\begin{aligned}&\max_{\{\theta_i\}} E\left[\widetilde{W}\right]-\frac{1}{2}\gamma Var\left[\widetilde{W}\right]\\ &\text{s.t.}\quad \widetilde{W}=W_0+\sum_{i=1}^{N_S}\theta_i\left(\widetilde{x}_i-P_i\right).\end{aligned} \tag{41}$$

Plugging in the budget constraint and the payoff $\widetilde{x}_i=a_i+b_i\widetilde{y}+\widetilde{e}_i$ of stock i, we can express the objective function as

$$\max_{\{\theta_i\}}\sum_{i=1}^{N_S}\theta_i\left(a_i-P_i\right)-\frac{1}{2}\gamma\left(\sum_{i=1}^{N_S}\theta_i b_i\right)^2\sigma_m^2-\frac{1}{2}\gamma\sum_{i=1}^{N_S}\theta_i^2\sigma_{e_i}^2. \tag{42}$$

The first-order condition with respect to θ_i is then given by

$$a_i-P_i-\gamma\left(\sum_{j=1}^{N_S}\theta_j b_j\right)b_i\sigma_m^2-\gamma\sigma_{e_i}^2\theta_i=0. \tag{43}$$

In equilibrium, the investor holds the available supply $\theta_i=u_i$ of stock i, which determines the stock price:

$$P_i=a_i-\gamma\left[\sigma_m^2\left(\sum_{j\neq i}u_j b_j\right)b_i+\left(\sigma_m^2 b_i^2+\sigma_{e_i}^2\right)u_i\right], \tag{44}$$

where we separated the terms that depend on the stock's own supply u_i.

A.2 Active Manager

The end investor's utility depends only on the Sharpe ratio of the market and the appraisal ratio of the active portfolio. Hence, the active manager's problem is to maximize the end investor's appraisal ratio, subject to the condition that the manager breaks even:

$$\begin{aligned}&\max_{\{\{v_j\},f\}}\frac{\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}|v_j|}{\sqrt{\sum_{j=1}^{N_S}v_j^2\sigma_j^2}}\\ &\text{s.t.}\quad f\frac{\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}|v_j|}{\gamma_e\sum_{j=1}^{N_S}v_j^2\sigma_j^2}\sum_{j=1}^{N_S}|v_j|\geq C.\end{aligned} \tag{45}$$

We write the Lagrangian of this problem as

$$\frac{\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}|v_j|}{\sqrt{\sum_{j=1}^{N_S}v_j^2\sigma_j^2}}+\lambda\left[f\frac{\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}|v_j|}{\gamma_e\sum_{j=1}^{N_S}v_j^2\sigma_j^2}\sum_{j=1}^{N_S}|v_j|-C\right]. \tag{46}$$

The first-order conditions with respect to v_i and f yield:

$$\frac{\left[\alpha_i-fs(v_i)\right]\sqrt{\sum_{j=1}^{N_S}v_j^2\sigma_j^2}-\left[\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}\left|v_j\right|\right]\left(\sum_{j=1}^{N_S}v_j^2\sigma_j^2\right)^{-\frac{1}{2}}\sigma_i^2v_i}{\sum_{j=1}^{N_S}v_j^2\sigma_j^2}+$$

$$+\frac{\lambda f}{\gamma_e}\left[\frac{\left[\left(\alpha_i-fs(v_i)\right)\left(\sum_{j=1}^{N_S}\left|v_j\right|\right)+\left(\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}\left|v_j\right|\right)s(v_i)\right]\sum_{j=1}^{N_S}v_j^2\sigma_j^2-\left(\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}\left|v_j\right|\right)\left(\sum_{j=1}^{N_S}\left|v_j\right|\right)2\sigma_i^2v_i}{\left(\sum_{j=1}^{N_S}v_j^2\sigma_j^2\right)^2}\right]=0$$

$$-\frac{\sum_{j=1}^{N_S}\left|v_j\right|}{\sqrt{\sum_{j=1}^{N_S}v_j^2\sigma_j^2}}+\frac{\lambda}{\gamma_e}\left[\frac{\left(\sum_{j=1}^{N_S}v_j\alpha_j-f\sum_{j=1}^{N_S}\left|v_j\right|\right)\left(\sum_{j=1}^{N_S}\left|v_j\right|\right)}{\sum_{j=1}^{N_S}v_j^2\sigma_j^2}-\frac{f\left(\sum_{j=1}^{N_S}\left|v_j\right|\right)^2}{\sum_{j=1}^{N_S}v_j^2\sigma_j^2}\right]=0$$

where $s(v_i)$ indicates the sign of v_i. We take the term containing λ from the latter first-order condition and substitute it into the former, which allows us to get rid of the sign functions. This gives us the portfolio weights:

$$v_i = \frac{\alpha_i}{\sigma_i^2} \frac{\left(1+\frac{\lambda f}{\gamma}\right)\left(\sum_{j=1}^{N_S} v_j^2 \sigma_j^2\right)}{\left(\sum_{j=1}^{N_S} v_j \alpha_j - f \sum_{j=1}^{N_S} |v_j|\right)\left(\sqrt{\sum_{j=1}^{N_S} v_j^2 \sigma_j^2} + 2\frac{\lambda f}{\gamma} \sum_{j=1}^{N_S} |v_j|\right)}. \tag{47}$$

The weights are thus proportional to $\frac{\alpha_i}{\sigma_i^2}$. We normalize the portfolio by requiring that $\sum_{v_i>0} v_i = 1$, which produces the final expression for the portfolio weights:

$$v_i = \frac{1}{\sum_{\alpha_j>0} \frac{\alpha_j}{\sigma_j^2}} \frac{\alpha_i}{\sigma_i^2}. \tag{48}$$

A.3 Analogy with a CARA Investor

Modeling the active manager as a CARA investor with a coefficient of absolute risk aversion γ, we let him solve the following maximization problem:

$$\begin{aligned} &\max_{\{W_i\}} E\left[-\exp\left(-\gamma \widetilde{W}_{a,1}\right)\right] \\ &\text{s.t. } \widetilde{W}_{a,1} = W_a + \sum_{i=1}^{N_S} W_i \widetilde{R}_i \end{aligned} \tag{49}$$

where $\widetilde{R}_i = \alpha_i + \beta_i \widetilde{R}_m + \widetilde{\varepsilon}_i$ is the excess return on stock i. Without loss of generality, we construct N_S uncorrelated hybrid securities $\widetilde{z}_i$ where each such security consists of one unit of a stock and a market hedge. The payoff of security i will then be $\widetilde{z}_i = a_i + \widetilde{e}_i$, and its price today will be $P_{z_i} = P_i + b_i \eta$. We then express the active manager's problem as:

$$\begin{aligned} &\max_{\{W_i\}} E\left[-\exp\left(-\gamma \widetilde{W}_{a,1}\right)\right] \\ &\text{s.t. } \widetilde{W}_{a,1} = W_a + \sum_{i=1}^{N_S} W_i \widetilde{R}_{z_i} + W_m \widetilde{R}_m \end{aligned} \tag{50}$$

where

$$\widetilde{R}_{z_i} = \frac{\widetilde{z}_i}{P_{z_i}} - 1 = \frac{a_i + \widetilde{e}_i}{P_{z_i}} - 1 = \frac{a_i + \widetilde{e}_i - (P_i + b_i \eta)}{P_i + b_i \eta} \tag{51}$$

is the excess return on hybrid security i. Since the payoffs of the hybrid securities are independent, the CARA investor will have a dollar demand for security i of

$$W_{z_i} = \frac{E\left[\widetilde{R}_{z_i}\right]}{\gamma Var\left[\widetilde{R}_{z_i}\right]} = \frac{1}{\gamma}\left[\frac{a_i - (P_i + b_i\eta)}{P_i + b_i \eta}\right]\left[\frac{(P_i + b_i\eta)^2}{\sigma_{e_i}^2}\right] = \frac{(a_i - b_i\eta - P_i)(P_i + b_i\eta)}{\gamma \sigma_{e_i}^2}. \tag{52}$$

As each hybrid security i consists of one share of stock i, the implied dollar demand for stock i is then

$$W_i = W_{z_i}\frac{P_i}{P_{z_i}} = \frac{(a_i - b_i\eta - P_i)(P_i + b_i\eta)}{\gamma\sigma_{e_i}^2}\frac{P_i}{(P_i + b_i\eta)} = \frac{(a_i - b_i\eta - P_i)P_i}{\gamma\sigma_{e_i}^2}. \tag{53}$$

To obtain a more intuitive expression, we substitute in the abnormal return of the stock (α_i) and the idiosyncratic variance of stock return (σ_i^2; not to be confused with payoff variance $\sigma_{e_i}^2$):

$$W_i = \frac{1}{\gamma}\left[\frac{a_i - b_i\eta - P_i}{P_i}\right]\left[\frac{P_i^2}{\sigma_{e_i}^2}\right] = \frac{\alpha_i}{\gamma\sigma_i^2}. \tag{54}$$

Note that each position in a hybrid security i will also generate a dollar demand of b_i for the market portfolio (to hedge market risk) and a dollar demand of $-b_i(1+\eta)$ for the risk-free asset. In equilibrium it will turn out that these hedging demands from the long and short positions perfectly cancel out as the active manager holds symmetric share positions around zero, so we do not need to address the question of whether the active manager should hedge market risk of the stock positions on his own or leave it to the end investor.

An unconstrained CARA investor would also have a “speculative” dollar demand of

$$W_m = \frac{\eta}{\gamma\sigma_m^2} \tag{55}$$

for the market portfolio directly. We set this demand equal to zero because the end investor should not reward the active manager for investing in the market portfolio. In the previous optimization problem (45) of the manager we did the same thing implicitly as we considered only abnormal returns and the market portfolio of course has an abnormal return of zero.

A.4 Equilibrium

We denote the supply of stock i left to the active manager as u_i. For the market to clear, the dollar supply has to equal the dollar demand, and this gives us the stock price:

$$\begin{aligned} u_i P_i &= W_i = \frac{\alpha_i}{\gamma\sigma_i^2} = \frac{(a_i - b_i\eta - P_i)\, P_i}{\gamma\sigma_{e_i}^2} && (56) \\ \Rightarrow \quad P_i &= a_i - b_i\eta - \gamma\sigma_{e_i}^2 u_i. && (57) \end{aligned}$$

The alpha of the stock is then:

$$\alpha_i = \frac{E\left[\widetilde{x}_i\right]}{P_i} - \beta_i\eta - 1 = \frac{a_i - b_i\eta - P_i}{P_i} = \frac{\gamma\sigma_{e_i}^2 u_i}{P_i}. \tag{58}$$

The market portfolio has an alpha of zero by construction. Hence,

$$\begin{aligned} \alpha_m &= \frac{\sum_{i=1}^{N_S} P_i\alpha_i}{\sum_{i=1}^{N_S} P_i} = 0 && (59) \\ \Rightarrow \quad \sum_{i=1}^{N_S} P_i\alpha_i &= \sum_{i=1}^{N_S} P_i \frac{\gamma\sigma_{e_i}^2 u_i}{P_i} = \gamma\sigma_e^2 \sum_{i=1}^{N_S} u_i = 0 && (60) \\ \Rightarrow \quad \sum_{i=1}^{N_S} u_i &= 0. && (61) \end{aligned}$$

Here we denote $\sigma_e^2 = E\left[\sigma_{e_i}^2\right]$ as the average idiosyncratic payoff variance across all stocks and assume $\sigma_{e_i}^2$ is uncorrelated with u_i. Since $u_{\bullet} + u_{in} + u_i = 1$, and since the passive manager’s position $u_{\bullet}$ is constant across stocks while the noise trader’s position u_{in} is distributed as $N\left(0, \sigma_u^2\right)$, the above equation implies the same distribution for the active manager’s equilibrium share holdings u_i:

$$u_i \sim N\left(0, \sigma_u^2\right). \tag{62}$$

As the noise trader and the active manager hold an average of zero of each stock, the passive manager has to hold the entire supply of 1 share, and hence he will hold the entire market portfolio:

$$u_{\bullet} = \frac{W_{\bullet}}{P_m} = 1. \tag{63}$$

Denoting the price of the market portfolio as P_m, its expected payoff as a_m, and the dollar variance of that payoff as σ_M^2, and plugging in the end investor’s allocation to the passive manager, we obtain the equilibrium market risk premium

$$\begin{aligned} 1 &= \frac{W_\bullet}{P_m} = \left(\frac{\eta}{\gamma_e \sigma_m^2}\right)\frac{1}{P_m} = \left(\frac{\eta P_m^2}{\gamma_e \sigma_M^2}\right)\frac{1}{P_m} = \frac{\eta a_m}{\gamma_e \sigma_M^2 (1+\eta)} && (64) \\ \Rightarrow \quad \eta &= \frac{\gamma_e \sigma_M^2}{a_m - \gamma_e \sigma_M^2} && (65) \end{aligned}$$

and the equilibrium allocation to the passive manager

$$W_\bullet = P_m = \frac{a_m}{1+\eta} = a_m - \gamma_e \sigma_M^2. \tag{66}$$

To find out the allocation to the active manager, we need to find the before-fee alpha of the manager:

$$\alpha_{bf} = \frac{\sum_{i=1}^{N_S} P_i u_i \alpha_i}{\sum_{u_i>0} P_i u_i}. \tag{67}$$

The cost of the portfolio is determined by the long positions, so only the long positions show up in the denominator. The numerator can be expressed as

$$\sum_{i=1}^{N_S} P_i u_i \alpha_i = \sum_{i=1}^{N_S} \gamma \sigma_{e_i}^2 u_i^2 = \gamma \sigma_e^2 \sum_{i=1}^{N_S} u_i^2 = \gamma \sigma_e^2 N_S \sigma_u^2. \tag{68}$$

For this aggregation, we used the assumption that there is a continuum of stocks with a measure of N_S, so $\frac{1}{N_S}\sum_{i=1}^{N_S} u_i^2 = E\left[u_i^2\right] = \sigma_u^2$. If we do not make the assumption, our results will be in the terms of particular realizations of all the u_i's (N_S of them), so the increase in mathematical rigor would come at the high cost of eliminating the simplicity and transparency of the equilibrium expressions. Due to the law of large numbers, this approximation does not affect our results in any meaningful way. Similarly for the denominator, we get

$$\begin{aligned} \sum_{u_i>0} P_i u_i &= \sum_{u_i>0} \left(a_i - b_i \eta - \gamma \sigma_{e_i}^2 u_i\right) u_i = (a - b\eta) \sum_{u_i>0} u_i - \gamma \sigma_e^2 \sum_{u_i>0} u_i^2 \\ &= \frac{N_S \sigma_u}{2}\left[\sqrt{\frac{2}{\pi}}(a - b\eta) - \gamma \sigma_e^2 \sigma_u\right] && (69) \end{aligned}$$

We also need the idiosyncratic variance of the active manager's portfolio. That is simply

$$\sigma_a^2 = \frac{\sum_{i=1}^{N_S} u_i^2 \sigma_{e_i}^2}{\left(\sum_{u_i>0} P_i u_i\right)^2} = \frac{4\sigma_e^2}{N_S\left[\sqrt{\frac{2}{\pi}}(a - b\eta) - \gamma \sigma_e^2 \sigma_u\right]^2}. \tag{70}$$

The fee of the active manager as a percentage of the cost of the portfolio is given by

$$f_a = \frac{\sum_{i=1}^{N_S} |W_i|}{\sum_{u_i>0} P_i u_i} f \tag{71}$$

where the numerator is

$$\sum_{i=1}^{N_S} |W_i| = \sum_{i=1}^{N_S} P_i |u_i| = -\sum_{u_i<0} P_i u_i + \sum_{u_i>0} P_i u_i = \sqrt{\frac{2}{\pi}}(a - b\eta) N_S \sigma_u, \tag{72}$$

and thus we get

$$f_a = \frac{2\sqrt{\frac{2}{\pi}}(a - b\eta) f}{\sqrt{\frac{2}{\pi}}(a - b\eta) - \gamma \sigma_e^2 \sigma_u}. \tag{73}$$

Finally, we can obtain the end investor's allocation to the active manager which depends on the after-fee alpha:

$$W_a = \frac{\alpha_{bf} - f_a}{\gamma_e \sigma_a^2} = \frac{N_S\left[-\sqrt{\frac{2}{\pi}}(a - b\eta) f + \gamma \sigma_e^2 \sigma_u\right]\left[\sqrt{\frac{2}{\pi}}(a - b\eta) - \gamma \sigma_e^2 \sigma_u\right]}{2\gamma_e \sigma_e^2}. \tag{74}$$

Equating this with the cost of the manager's portfolio (69), we obtain the simple formula for the effective risk aversion of the manager:

$$\gamma = \gamma_e + \sqrt{\frac{2}{\pi}} \left(\frac{a - b\eta}{\sigma_e^2 \sigma_u} \right) f. \tag{75}$$

The value of the proportional fee f allows the manager to exactly cover his fixed dollar cost C:

$$\begin{aligned} C &= f \sum_{i=1}^{N_S} |W_i| = \sqrt{\frac{2}{\pi}} \left(a - b\eta\right) N_S \sigma_u f & (76) \\ \Rightarrow \quad f &= \frac{C}{\sqrt{\frac{2}{\pi}} \left(a - b\eta\right) N_S \sigma_u}. & (77) \end{aligned}$$

A.5 Welfare in Alternative Market Designs

A.5.1 Certainty Equivalent of Zero

Note that in the market designs discussed before (including the CAPM), investors actually had a positive certainty equivalent for holding the active portfolio. Hypothetically we could squeeze them even further, requiring that their certainty equivalent is zero, which translates to the following problem:

$$\begin{aligned} & \max_{P, \gamma_c} \quad P \\ \text{s.t.} \quad & \left(E\left[\widetilde{X}\right] - P \right) - \frac{1}{2} \gamma_c Var\left[\widetilde{X}\right] \geq 0 \\ & \gamma_c \geq \gamma_e. \end{aligned} \tag{78}$$

The welfare-maximizing price is then

$$P = -C - \frac{1}{2} \gamma_e N_S \sigma_u^2 \sigma_e^2. \tag{79}$$

Here we actually get slightly flatter demand curves than before, because we no longer require that the investors can choose their own exposure to the active portfolio. The implementation of this result requires that each investor is given a personal (non-transferable) offer to invest a fixed dollar amount in the active portfolio for a fixed fee. The dollar amounts of the offer (both investment and fee) are tailored to the risk aversion of each individual investor to leave each of them exactly indifferent between investing and not investing. A secondary market in such claims has to be banned.

Comparing the value of P in this extreme situation with the one we had before (equation (37)), an interesting result emerges. In our calibration with a fee of 1.5%, the cost C is about 1,000 times greater than the term $\gamma_e N_S \sigma_u^2 \sigma_e^2$ which involves end investors' risk aversion. Hence, the numerical difference between the two results is negligible. This is perhaps surprising, given the extreme difficulty of implementing (79) and the great simplicity of implementing (37). We can therefore conclude that for any practical purposes our institutional structure is indeed optimal also across all hypothetical market designs.

A.5.2 Example of an Alternative Market Design

For comparison, consider an alternative market design with a fixed cost C but without institutions. Then we will be left with the basic CAPM setting, except that some investors pay the fixed cost and share it among one another (assuming they are all identical and they share the cost per head), while others do not pay anything and only invest in the market portfolio.

Since prices are set by CARA investors, the price of stock i will still be equal to $P_i = a - b\eta - \gamma\sigma_{e_i}^2 u_i$ where γ is the collective risk aversion of those investors who have paid the cost. In equilibrium, the fraction μ of the investors who pay the cost has to be such that an investor is indifferent between paying and not paying. It turns out to be

$$\mu = \frac{\gamma_e N_S \sigma_e^2 \sigma_u^2}{2C}, \tag{80}$$

which generally represents a very small fraction (about 0.0005 if the fee is 1.5%). This means that the risk aversion of those active investors collectively will be

$$\gamma = \max\left\{\gamma_e, \frac{2C}{N_S \sigma_e^2 \sigma_u^2}\right\}. \tag{81}$$

In our model with institutions, the effective risk aversion of the active manager was instead

$$\gamma = \gamma_e + \frac{C}{N_S \sigma_e^2 \sigma_u^2}, \tag{82}$$

where the first term (γ_e) in the summation is negligible for interesting values of the cost (and the proportional fee). Hence, demand curves in this setting will be twice as steep as in the presence of institutions.

In this setting without institutions, the idiosyncratic risk of the active portfolio is borne by only a small subset of investors and thus they require greater risk premium for it. In contrast, institutions allow every investor in the economy to bear a small fraction of the active portfolio, which results in more efficient risk-sharing and thus flatter demand curves.

A.6 CRRA Utility

Our standard CAPM calibration assumes normally distributed asset payoffs and CARA utility. To confirm the robustness of our results to the utility specification, we also consider a representative investor with CRRA utility and a coefficient of relative risk aversion of γ_R. The local coefficient of absolute risk aversion of the investor is then $\widetilde{\gamma} = \frac{\gamma_R}{\widetilde{W}}$, where $\widetilde{W}$ is the wealth of the investor. Since the wealth of the investor depends on the random return on the market portfolio, also the local coefficient of absolute risk aversion will be random. We assume a lognormal distribution for the market return which implies a lognormal distribution for $\widetilde{\gamma}$.

To work out a tractable solution, we need two approximations. First, because any idiosyncratic gamble will be a negligible part of the investor's wealth (the market portfolio has 4×10^7 times the dollar variance of 10% supply of a stock and an expected payoff of about 10^4 times as much), we approximate the CRRA investor as "locally CARA" where the coefficient of absolute risk aversion depends only on the random return on the market portfolio. Second, we approximate the lognormal distribution of $\widetilde{\gamma}$ with a normal distribution.

Consider a normally distributed idiosyncratic net payoff $\widetilde{Z} = \theta\left(\widetilde{x} - P\right)$. The dollar payoff of one share is $\widetilde{x}$, and it has a price of P. The investor buys θ units of it. We can then derive the investor's demand for the payoff by solving:

$$\begin{aligned} &\max_{\{\theta\}} E\left[-\exp\left(-\widetilde{\gamma}\widetilde{Z}\right)\right] \\ &\text{s.t.} \quad \widetilde{Z} = \theta\left(\widetilde{x} - P\right). \end{aligned} \tag{83}$$

Of course now the standard mean-variance analysis will not go through since $\widetilde{\gamma}$ is not constant, and the objective function does not simplify to the usual mean-variance problem. Instead, using iterated expectations we obtain:

$$E\left[-\exp\left(-\widetilde{\gamma}\widetilde{Z}\right)\right] = E\left[E\left[-\exp\left(-\widetilde{\gamma}\widetilde{Z}\right)|\widetilde{\gamma}\right]\right] = E\left[-\exp\left(-\widetilde{\gamma}E\left[\widetilde{Z}\right] + \frac{1}{2}\widetilde{\gamma}^2 Var\left[\widetilde{Z}\right]\right)\right]. \tag{84}$$

For notational simplicity, we write $\mu_Z = E\left[\widetilde{Z}\right]$, $\sigma_Z^2 = Var\left[\widetilde{Z}\right]$, $\mu_\gamma = E\left[\widetilde{\gamma}\right]$, $\sigma_\gamma^2 = Var\left[\widetilde{\gamma}\right]$. The above integral can be evaluated if $\sigma_\gamma^2\sigma_Z^2 \leq 1$, a condition that is easily satisfied. After some algebra, the maximization problem simplifies to

$$\max_{\{\theta\}} \frac{-1}{\sqrt{1-\sigma_\gamma^2\sigma_Z^2}} \exp\left[\frac{1}{2\sigma_\gamma^2} \times \frac{\left(\sigma_\gamma^2\mu_Z - \mu_\gamma\right)^2}{1-\sigma_\gamma^2\sigma_Z^2}\right]. \tag{85}$$

Plugging in $\mu_Z = \theta\left(E\left[\widetilde{x}\right] - P\right)$ and $\sigma_Z^2 = \theta^2 Var\left[\widetilde{x}\right]$, the first-order condition gives us

$$\begin{aligned} \theta Var\left[\widetilde{x}\right]\left[\sigma_\gamma^2 + \left(\sigma_\gamma^2\theta\left(E\left[\widetilde{x}\right] - P\right) - \mu_\gamma\right)^2\right] &= \left[\mu_\gamma - \sigma_\gamma^2\theta\left(E\left[\widetilde{x}\right] - P\right)\right]\left(E\left[\widetilde{x}\right] - P\right) & (86) \\ \Rightarrow \quad \theta Var\left[\widetilde{x}\right]\left[\sigma_\gamma^2 + \mu_\gamma^2\right] &= \mu_\gamma\left(E\left[\widetilde{x}\right] - P\right), & (87) \end{aligned}$$

where we ignored the insignificant terms in the last line. Hence, we can write the approximate demand of the CRRA investor as:

$$\theta_{CRRA} = \frac{E\left[\widetilde{x}\right] - P}{Var\left[\widetilde{x}\right]} \times \frac{1}{\mu_\gamma + \frac{\sigma_\gamma^2}{\mu_\gamma}}. \tag{88}$$

Note the similarity to the demand of a standard CARA investor with a constant coefficient of absolute risk aversion μ_γ:

$$\theta_{CARA} = \frac{E\left[\widetilde{x}\right] - P}{Var\left[\widetilde{x}\right]} \times \frac{1}{\mu_\gamma}. \tag{89}$$

In the CAPM calibration with a one-year horizon, we have approximately $\sigma_\gamma = 0.2\mu_\gamma$, so $\theta_{CRRA} = \frac{\theta_{CARA}}{1.04}$. This means a 4% reduction in demand which in turn implies a 4% increase in the slope of the demand curve, or a 4% increase in price impact. Starting from the tiny CAPM price impact of 0.11 bp for a 10% supply shock, the increase of 0.004 bp is clearly meaningless. For a 5-year horizon we have $\sigma_\gamma = 0.45\mu_\gamma$ and $\theta_{CRRA} = \frac{\theta_{CARA}}{1.20}$, which would increase the 0.62 bp price impact to 0.74 bp. Again, there is no meaningful difference between the CARA and CRRA cases.

In our model with institutions, the answer is even simpler because it turned out that the risk aversion of the end investors has a negligible impact on the slopes of demand curves. The equilibrium slope of the demand curve will be set so that the active manager's net-of-fee alpha is just barely above zero. The end investors' demand and thus the net-of-fee alpha would have to be scaled up not by some percent but by several orders of magnitude before it has any pricing implications. Hence, there seems to be no reason why the economic results of this paper would be sensitive to the convenient utility specification.

B Figures

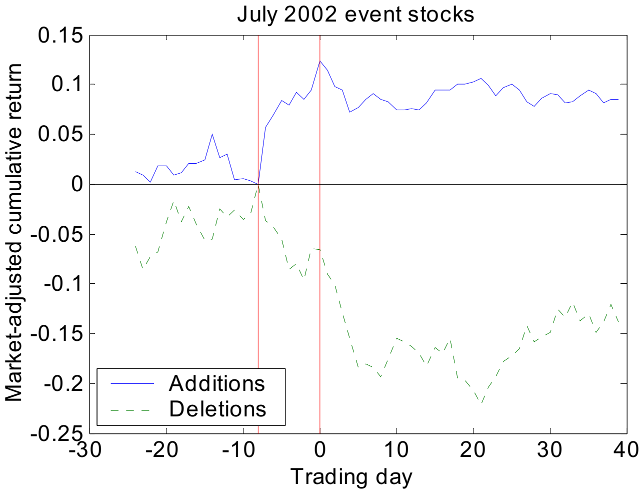

Figure 3: July 2002 replacement of seven non-U.S. firms in the S&P 500 index. The announcement occurred after the close on trading day −8, while the changes became effective at the close on trading day 0. The graph shows buy-and-hold returns on portfolios formed (initially with equal weights) on trading day −8.

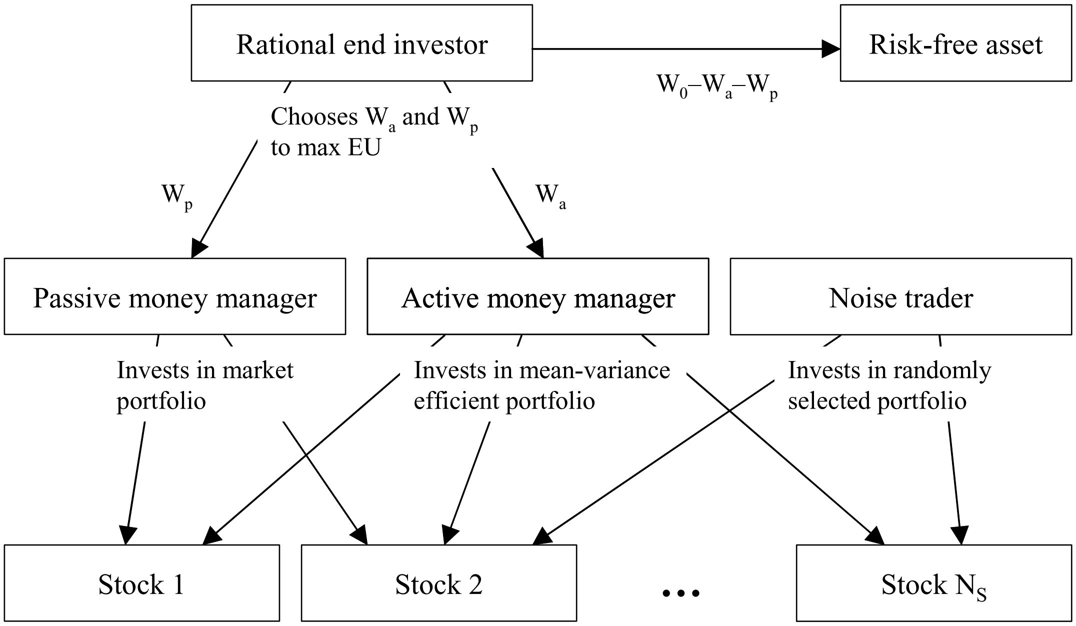

Figure 4: The basic setup for the model.

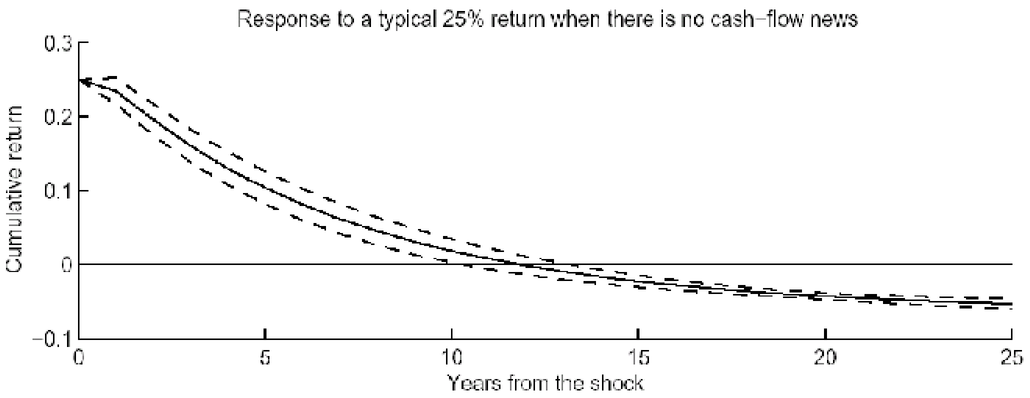

Figure 5: Speed of reversal of an expected-return shock, taken from Cohen, Gompers, and Vuolteenaho (2002).

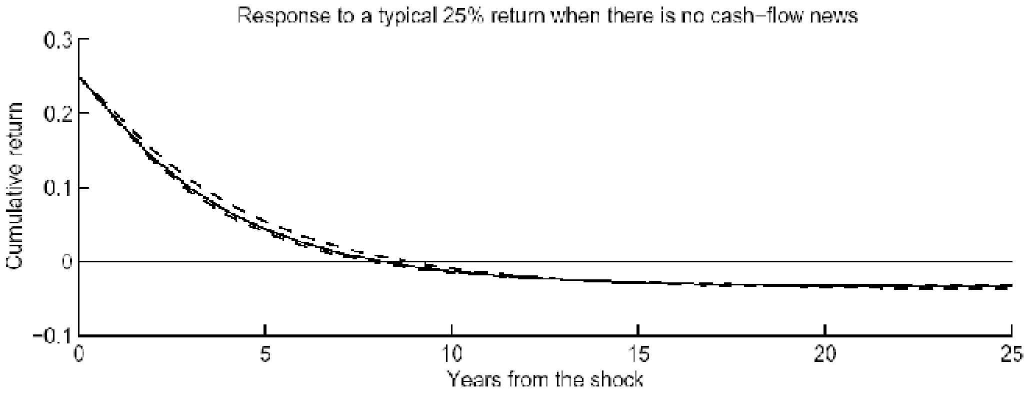

Figure 6: Speed of reversal of an expected-return shock, taken from Cohen, Gompers, and Vuolteenaho (2002).