Do Intermediaries Matter for Stock Returns? A Dynamic Demand System Approach*

Di Tian

University of Pennsylvania

This Version: February 21, 2024

Abstract

I develop a two-tier asset demand system that incorporates endogenous aggregate allocation and short sales, and propose a two-step estimation procedure with a novel instrument for aggregate estimation, which allows me to exploit both cross-sectional and time-series variation in institutional holdings. The estimated system provides a framework to answer questions related to demand-side effects of financial intermediaries and short sales in both aggregate and individual stock markets. I find institutional demand accounts for a large proportion, if not all, of observed return premiums in size, value and investment. The short leg, while increasingly important, cannot explain observed anomaly returns and the formation of the dot-com bubble. However, short sales do have significant yet disparate pricing impact on stocks with different characteristics. In the aggregate stock market, unobserved aggregate preference and beliefs rather than risk-return balance is the main driver of the return predictability of dividend-price ratio.

*I thank my committee, Winston Wei Dou, Xu Cheng, and Karun Adusumilli for their guidance and support. Additionally, I thank Andrew Abel, Luke Taylor, Vincent Glode, Sean Myers, Frank Diebold, Frank Schorfheide, and seminar participants at Wharton and Penn Economics for comments.

1 Introduction

Recent decades have witnessed the increasingly dominating role of financial intermediaries in the United States financial markets. Institutional ownership share of equity has grown from 34% to 67% since 1980 as the number of institutions has more than quintupled. Abundant studies have documented the importance of intermediaries to asset prices, most of which offer some insight with theories and reduced-form evidence (e.g., Basak and Pavlova (2013); Vayanos and Woolley (2013); Adrian et al. (2014); Frazzini and Pedersen (2014); He et al. (2017); Drechsler et al. (2018); Dou et al. (Forthcoming)). However, to quantitatively examine narratives about demand-side effects of institutions, a demand system with market clearing of all investors is needed.

Koijen and Yogo (2019) are the first to propose a characteristics-based demand system that matches institutional holdings data. They provide a framework to study the asset pricing impact of institutional demand in a structural way, though they simplify several key aspects. First, they ignore the endogenous choice between risk-free bonds and risky assets and assume the aggregate flow into the stock market is fixed. This might not be a reasonable assumption when studying demand-side effects due to market-wide changes. Related, their framework only exploits cross-sectional variation in holdings data and disregards time-series variation, which is necessary for aggregate estimation. Finally, they neglect short sales, which are a unique and important aspect of institutional demand. Plenty of evidence show their relevance to stock returns, though their exact role is still a matter of heated debate (e.g., Stambaugh et al. (2012); Geczy et al. (2002); Beber and Pagano (2013)).

To answer questions about institutional demand and short selling, I propose a two-tier demand system that incorporates both endogenous aggregate allocation and short sales, inspired by the nested logit model from the differentiated products demand literature (Rosen (1974); McFadden (1981); Cardell (1997); Gandhi and Nevo (2021)).

Specifically, in the first tier, each investor distributes their resources between a risk-free bond and the risky asset class (which covers mostly equity securities in the data), depending on the market conditions, aggregate latent demand, and the diversification within the risky asset class. This diversification is derived from the second-tier demand, in which investors form their sub-portfolios across individual risky assets based on their preference over asset characteristics and residual latent demand. The nested logit model decomposes the total unobserved demand for individual risky assets into a common aggregate component and a residual asset-specific component, which are assumed to be orthogonal to each other. This allows me to incorporate time-series co-movement among demand for individual risky assets,

which enriches the demand structure. In the meantime, it allows me to easily separate the demand into two tiers, which simplifies the estimation. Overall, the adoption of the nested logit model emphasizes the distinct difference between risk-free bonds and risky assets and allows for a more flexible substitution pattern across assets.

I match the demand system with the detailed institutional holdings data obtained from Securities and Exchange Commission (SEC) Form 13F. Form 13F is a quarterly institutional holdings report that is required to be filed by all institutional investment managers with at least $100 million in assets under management since 1980. It only discloses information on long positions, and data on institution-level short positions are unavailable. Instead, I utilize data on stock-level short interest from Compustat to construct a representative short seller in order to investigate the asset pricing effects of short sales.

The identification strategy follows a two-step procedure. First, I estimate the second-tier demand within the risky asset class. Conditioning on total investment in risky assets, the conditional weights of individual risky assets are solely determined by asset characteristics and asset-specific latent demand. Following the traditional practice, I assume shares outstanding and characteristics other than price are exogenous. To address the endogeneity of price, Koijen and Yogo (2019) construct an instrument by extracting exogenous components of demand from investors' investment universe and total risky capital. In my model, however, asset-specific latent demand also affects how much capital is allocated to the risky asset class, endogenizing current investment in risky assets. Therefore, I improve their instrument by replacing current investment in risky assets with average past investment in risky assets, which extracts the low-frequency component that is reasonably exogenous to current latent demand.

In the second step, the estimation of aggregate demand poses two major challenges. Most importantly, due to the absence of risk-free holdings data, the above approach cannot be implemented in the aggregate tier. Instead, I adopt the log-linearization technique from Gabaix and Koijen (2021), so that only data on risky asset holdings are necessary. The estimation uses the dynamic panel of all investors, which allows me to exploit the time-series aspect of the data disregarded by Koijen and Yogo (2019). The other challenge relates to the endogeneity of price and Sharpe ratio with the aggregate latent demand. I simply utilize the estimated asset-specific latent demand from the first step, which is imposed by the model to be orthogonal to the aggregate latent demand. The idea is that asset-specific latent demand, especially that of investors with larger market shares and assets with larger market capitalization, sufficiently affects asset prices and aggregate Sharpe ratio in equilibrium but

remain orthogonal to aggregate latent demand, making them valid instruments.

Once the two-tier demand system is estimated, I demonstrate the asset pricing impact of institutional demand with four empirical applications. First, I decompose the cross-sectional variance of stock returns into supply- and demand-side effects. The supply-side effects, including changes in shares outstanding, characteristics and dividend yield, only explains 10.7 percent of return variation. In sharp contrast, the demand-side effects, including changes in total wealth, aggregate preference and individual stock preference, respectively account for 2.5, 0.6 and 86.2 percent. Among these, asset-specific latent demand, which explains 83 percent, is the predominant source of return volatility, consistent with the findings of Koijen and Yogo (2019). Though aggregate latent demand merely accounts for 0.5 percent of return variation in the full sample, its effect more than doubled in the latter half of the sample, indicating a more flexible demand in aggregate equities.

Second, I inspect the connection between institutional demand and common return anomalies. When financial intermediaries no longer demand the corresponding characteristics, size and investment premiums become minuscule and insignificant, suggesting these anomalies can be mainly attributed to institutional demand. In contrast, value premium, though decreased by approximately 40 percent, stays significant, and profitability premium is virtually unaffected, suggesting they are more likely to be driven by fundamental reasons. On the other hand, several papers have focused on the short leg of returns (e.g., Stambaugh et al. (2012); Avramov et al. (2013); Drechsler and Drechsler (2014)), but I find no evidence that short-sale constraints drive the observed anomaly returns.

Third, I examine the role of short sales in stock valuation. I first decompose the cross-sectional variance of stock returns into long- and short-side effects. Results show that the effect of short sales first became significant around the dot-com bubble and has been growing to explain approximately 3–4 percent of return variation. This lends evidence to short sellers' role in unearthing over-valued stocks and their importance during financially abnormal times. I then examine the exact effect of short sales on stock prices, especially during the dot-com bubble and the 2008 financial crisis. I find that short-sale constraints were not crucial to the formation of the dot-com bubble, as the pricing impact of short sales were within normal range. During the financial crisis, a short-sale ban would significantly inflate stock prices, especially for large and less profitable firms but not for small firms. This asymmetric mispricing effect is also reflected in Daniel et al. (2022) who focus on different characteristics such as momentum. These findings could offer some insight into the debate on the effectiveness of short-sale policies during an economic bubble (Ofek and Richardson (2003); Battalio

and Schultz (2006)) or a financial crisis (Battalio et al. (2012); Boehmer et al. (2013)).

Finally, I use the model to investigate the underlying mechanism behind the phenomenon that dividend-price ratio forecasts future returns in the aggregate stock market, first documented by Campbell and Shiller (1988b). Using simulated demand, I isolate out the effect of risk-return balance and that of unobserved preference in the aggregate demand. When investors are neutral towards aggregate Sharpe ratio, the return predictability of dividend-price ratio weakens marginally. In contrast, when aggregate latent demand is substituted with idiosyncratic shocks, dividend-price ratio no long predicts future returns. It is even more so when both effects are combined. In conclusion, while risk aversion certainly plays a role (Campbell and Cochrane (1999)), unobserved preference and beliefs about the aggregate stock market, which nests the theory of Lakonishok et al. (1994), are more likely to be the main driver of the return predictability of dividend-price ratio.

Related Literature This paper mainly contributes to the literature on asset demand system. This strand of literature has roots going back to (at least) Brainard and Tobin (1968) and Tobin (1969). Tobin (1969) emphasizes the importance of interdependencies across financial markets, which is represented in the aggregate tier of my model. A later improvement, the Almost Ideal Demand System developed by Deaton and Muellbauer (1980), extends the model to portfolio choice and characteristics other than returns. However, its application to portfolio choice has been much less common than in consumption studies. Recently, due to better availability of holdings data, Koijen and Yogo (2019) have revived this strand of literature by estimating a characteristics-based demand system in the U.S. stock market. Demand systems for other asset markets or countries have also been considered, including U.S. Treasury bonds (Krishnamurthy and Vissing-Jorgensen (2007)), U.S. and U.K. stocks (Koijen et al. (2019)), euro-area government bonds (Koijen et al. (2021)), and aggregate global assets (Koijen and Yogo (2020); Gabaix et al. (2022)). This paper extends the framework of Koijen and Yogo (2019) in at least the following aspects. (i) I jointly model the demand for asset classes and individual assets, which endogenizes aggregate allocation and allows for more flexible substitution patterns within and across asset classes. (ii) I propose a novel instrument for aggregate estimation and exploit both cross-sectional and time-series variation. (iii) I quantitatively examine demand-side effects on movements in both aggregate and individual markets. For example, I investigate the asset pricing impact of aggregate demand and find that unobserved preference, rather than risk-return trade-off, is the main driver of the return predictability of dividend-price ratio in the aggregate stock market. (iv) I incorporate short sellers in the demand system and demonstrate the role of

short sales in a structural way that accounts for the interaction between the long side and the short side, which could shed some light on the debates about short-sale policies.

The methodology contribution is connected to the broad set of literature on empirical asset pricing methods beyond asset demand system. Many focus on the classical risk premium problem, including Fama and MacBeth (1973), Kan and Zhang (1999b), Kan and Zhang (1999a), Gu et al. (2020) and Giglio and Xiu (2021), while others look at the estimation and testing of other financial models, including Hansen et al. (1996), Aït-Sahalia and Yu (2006), Yu (2007) and Cheng et al. (2022). I refer to Campbell et al. (2012) for a more comprehensive review.

This paper is also related to the literature on the estimation of price elasticities in the stock market. Besides the demand system approach, many estimate micro-elasticities using index inclusion induced shocks (e.g., Shleifer (1986); Chang et al. (2015)), trade-level activities (e.g., Frazzini et al. (2018); Bouchaud et al. (2018)) or mutual fund flows (e.g., Lou (2012)). Other studies focus on macro-elasticities on the aggregate level (e.g., Da et al. (2018); Gabaix and Koijen (2021); Hartzmark and Solomon (2022)). I add to this literature by jointly evaluating micro- and macro-elasticities implied by the two-tier demand system.

Another relevant line of literature is on the the asset allocation of financial intermediaries (e.g., Daniel et al. (1997); Gompers and Metrick (2001); Basak et al. (2007); Cremers and Petajisto (2009); Hugonnier and Kaniel (2010); Cuoco and Kaniel (2011); Kacperczyk et al. (2014); Blume and Keim (2014); Pastor et al. (2020)) and their role in asset pricing (e.g., Goldman and Slezak (2003); Cornell and Roll (2005); Asquith et al. (2005); Kaniel and Kondor (2013); Basak and Pavlova (2013); Vayanos and Woolley (2013); Koijen (2014); He et al. (2017); Drechsler et al. (2018); Breugem and Buss (2018);Buffa and Hodor (2018); Gabaix and Koijen (2021); Dou et al. (Forthcoming)). These studies investigate the asset pricing implications of various aspects of intermediaries, including relative-performance-based compensation of managers, index-tracking restrictions, leverage ratio, and common flow risks. My paper also connects to the literature on the role of short sales with respect to return anomalies, bubble formation and stock valuation (e.g., Geczy et al. (2002); Ofek and Richardson (2003); Battalio and Schultz (2006); Haruvy and Noussair (2006); Stambaugh et al. (2012); Battalio et al. (2012); Boehmer et al. (2013); Avramov et al. (2013); Beber and Pagano (2013); Drechsler and Drechsler (2014); Daniel et al. (2022)). The primary shortcoming of these strands of literature is the difficulty of distinguishing which channel(s) are actually taking effect and how to quantify them due to the nature of reduced-form methods. By using a fully developed demand system, I provide a structural and quantitative frame-

work to examine how different aspects of institutional demand and short sales explain the observed phenomena in the stock market.

The remainder of this paper is organized as follows. Section 2 develops the two-tier asset demand system and discusses the implied demand elasticities and how market clearing determines asset prices. Section 3 describes the data on asset characteristics, aggregate Sharpe ratio and institutional holdings. Section 4 explains the identifying assumptions and estimation procedure and presents the estimates of the two-tier demand system. Section 5 illustrates the empirical relevance of the demand system and investigates the role of institutional demand and short sales. Finally, Section 6 concludes.

2 Two-Tier Asset Demand System

I propose a two-tier asset demand system, where long investors and a representative short seller allocate their capital among a risk-free bond and many risky assets. The portfolio demand employs the nested logit model of McFadden (1981) and Cardell (1997), which nests the characteristics-based demand of Koijen and Yogo (2019) as a special case. The weight of the risky asset class depends on the aggregate market conditions and the diversification value of risky assets, whereas the conditional weight of each individual risky asset within the class is determined by the investor's taste in asset characteristics.

2.1 Assets and Investors

In the entire investment universe, there are $N+1$ financial assets indexed by $n = 0, 1, ..., N$, which fall into two classes. One is the risk-free bond indexed by $n = 0$, and the other is the risky asset class, indexed by $n = *$, which covers assets $1 - N$. For individual asset n, denote P_t^n and Q_t^n as the price per share and the number of the shares outstanding at time t. ME_t^n is the corresponding market equity. The market equity of the aggregate risky asset market is defined as the total market value of all individual risky assets. Following the pricing methodology of many established stock market indices (e.g., S&P 500), I define the aggregate price of the risky asset class as

$$P_t^* = \frac{\sum_{n>0} P_t^n Q_t^n}{Divisor} \tag{1}$$

where the divisor is set to ensure a fixed supply in the aggregate risky asset market.

Let lowercase letters be the logarithm of the corresponding uppercase variables. That is, $p_t^n = \log(P_t^n)$, $q_t^n = \log(Q_t^n)$, and $me_t^n = \log(ME_t^n)$.

The financial assets are differentiated along two sets of characteristics. The first set, denoted as y_t^n, $n = 0, *$, includes the aggregate market conditions, which are commonly shared within an asset class. The second set, denoted as x_t^n, $n = 1, ..., N$, includes asset-specific characteristics for individual risky assets. In the case of stocks, for example, y_t^n could include aggregate Sharpe ratio, whereas x_t^n could include fundamentals such as market equity, book equity and profitability. Section 3 provides a detailed description of the characteristics. I denote the k-th element of the characteristics vectors as $y_{k,t}^n$ and $x_{k,t}^n$, and let $y_{0,t}^n$ and $x_{0,t}^n$ be the constant for convenience. Following the literature on asset pricing in endowment economies (Lucas (1978)), I assume that shares outstanding and characteristics other than aggregate Sharpe ratio and market equity are exogenous.

All assets are held by $I + 2$ investors. A representative specialized short seller, indexed by $i = -1$, performs short sales, while the other investors only hold long positions. Investors $i = 1, ..., I$ are financial institutions, grouped into six types including banks, insurance companies, investment advisors, mutual funds, pension funds, and other institutions (see Section 3 for more detail). The remaining assets are held by the household sector, indexed by $i = 0$.

2.2 Nested Logit Demand

Drawing inspiration from the nested logit model, I decompose asset demand into two tiers. For investor i at time t, in the first tier, they distribute their total wealth $A_{i,t}$ between different asset classes. Denote the total investments in risk-free bonds and risky assets as $A_{i,t}^0$ and $A_{i,t}^*$ respectively. In the second tier, the investor forms their conditional portfolio across individual risky assets that are in their investment universe $N_{i,t} \subset 1, ..., N$. Denote the dollar investment in individual risky asset n as $A_{i,t}^n$. This process is not necessarily executed sequentially. On the contrary, diversification within the second tier inversely affects first-tier demand as shown later. From the perspective of the discrete choice literature, each investor at each time can be viewed as a market and the portfolio weights can be viewed as the market shares of assets. The adoption of the nested logit model emphasizes the distinct difference between risk-free bonds and risky assets and allows for a more flexible substitution pattern across assets.

Note that the notion of investment universe $N_{i,t}$ is necessary because investors typically do not have full access to all risky assets either due to investment mandates, client restrictions

or management style. Here, I assume asset 1 is a base risky asset that all investors have access to, i.e., $1 \in N_{i,t}$. And let $|N_{i,t}|$ be the number of risky assets that are available to investor i at time t.

For the second-tier demand, I adopt the characteristics-based approach for portfolio choice (e.g., Lynch (2001); Brandt et al. (2009); Koijen and Yogo (2019)), in order to model individual asset differentiation and focus on estimating demand with a large number of assets. This approach directly models the portfolio weight of each risky asset as a function of asset characteristics. The basic idea is that asset characteristics are closely related to asset returns and variances and sufficiently capture the joint distribution. Compared to the traditional Markowitz and utility maximization approach, which are mainly theoretical frameworks involving a practically unmanageable number of higher moments and featuring unreliable guidance on real-time investment, the characteristics-based approach is likely a more relevant and stable approximation to the portfolio choice behavior in practice (Brandt and Santa-Clara (2006); Brandt et al. (2009)). In fact, Koijen and Yogo (2019) provide a heuristic argument to justify the approximate equivalence between the Markowitz approach and the characteristics-based approach. Therefore, following the characteristics-based approach, I model the relative weights among individual risky assets for investor i at time t as

$$\begin{aligned} \frac{w_{i,t}^n}{w_{i,t}^1} = \frac{w_{i,t}^{n|*}}{w_{i,t}^{1|*}} &= \delta_{i,t}^n, \qquad n \in N_{i,t} - \{1\} \\ &= \exp\left(\beta_{0,i,t} + \beta_{1,i,t} me_t^n + \sum_{k \geq 2} \beta_{k,i,t} x_{k,t}^n\right) \xi_{i,t}^n \end{aligned} \tag{2}$$

where $w_{i,t}^n = A_{i,t}^n / A_{i,t}$ is the actual portfolio weight of asset n, and $w_{i,t}^{n|*} = A_{i,t}^n / A_{i,t}^*$ is the conditional portfolio weight of asset n within the risky asset class. Equation 2 highlights that the relative demand for individual risky assets depends on asset characteristics x_t^n including price (log market equity) and asset-specific latent demand $\xi_{i,t}^n$. Let $\delta_{i,t}^1 = 1$ for notation convenience. Then, combined with the budget constraint, Equation 2 implies the conditional weights of risky asset n depend on observed characteristics and unobserved residual demand of all risky assets

$$w_{i,t}^{n|*} = \frac{\delta_{i,t}^{n}}{\sum_{n \in N_{i,t}} \delta_{i,t}^{n}} \tag{3}$$

These conditional weights are what Koijen and Yogo (2019) refer to as characteristics-based demand, which solely depend on asset-specific components and resemble the form of market shares in a logit model (Rosen (1974)). Therefore, the nested logit model nests the system of Koijen and Yogo (2019) as a special case and is a natural extension from the second-tier demand to the first-tier demand. It allows me to model the distinction between risk-free bonds and risky assets and to differentiate the substitution patterns within and across asset classes. The basic idea is that investors allocate their capital between asset classes based on the aggregate condition of each market and the overall investment value of individual assets within each market. Specifically, the relative portfolio weight of the risky asset class to the risk-free bond is

$$\frac{w_{i,t}^{*}}{w_{i,t}^{0}} = \delta_{i,t}^{*} = \exp\left(\alpha_{0,i,t} + \alpha_{1,i,t} SR_t + \sum_{k \geq 2} \alpha_{k,i,t} y_{k,t}^{*} + \theta \Gamma_{i,t}^{*} + \tilde{\xi}_{i,t}^{*}\right) \tag{4}$$

$$\Gamma_{i,t}^{*} = \log\left(\sum_{n \in N_{i,t}} \delta_{i,t}^{n}\right) \tag{5}$$

Equation 4 illustrates that the portfolio weight of the risky asset class depends on the market conditions y_t^* including aggregate Sharpe ratio, the composite value $\Gamma_{i,t}^{*}$ from the second-tier demand and the aggregate latent demand $\tilde{\xi}_{i,t}^{*}$. A key property of this model is the fact that the total latent demand of individual risky assets is decomposed into a common aggregate component $\tilde{\xi}_{i,t}^{*}$ in Equation 4 and a residual asset-specific component $\xi_{i,t}^{n}$ in Equation 2. The two are assumed to be orthogonal to each other in the model, which is one of the key identification assumptions utilized by the estimation procedure discussed in Section 4. This decomposition highlights the advantages of the nested logit model by allowing co-movement among demand for individual risky assets while also separating the two-tier demand, which enriches the structure and simplifies the estimation.

The value $\Gamma_{i,t}^{*}$, which I call the diversification value of the risky asset class, is an important contribution of the nested logit model. It tightly connects the first-tier demand and the second-tier demand and the coefficient $\theta \in (0, 1)$ governs the substitution pattern across asset classes. Mathematically, $\Gamma_{i,t}^{*}$ is the "inclusive value" of the risky asset class in the nested logit

model, which measures the expected maximum utility attained from risky assets and links portfolio weights within and outside of an asset class. Intuitively, it is a composite index of demand for individual risky assets, which captures the gains from strategically diversifying investments within the risky asset class according to the investor's personal preference over different characteristics. It closely relates to measures of portfolio diversification in previous literature. Goetzmann and Kumar (2008), Ivkovic et al. (2008) and Pollet and Wilson (2008) use the number of stocks in the portfolio to measure diversification, while Blume and Friend (1975) and Pastor et al. (2020) use the deviation of portfolio weights from the market. The nested logit measure of diversification incorporates both aspects, because the number of available assets and deviation from the market to suit the investor's preference both increase this value.

Equation 4, combined with the budget constraint, gives the portfolio weight of each asset class at time t

$$w_{i,t}^{*} = \frac{\delta_{i,t}^{*}}{1+\delta_{i,t}^{*}}, \qquad w_{i,t}^{0} = \frac{1}{1+\delta_{i,t}^{*}} \tag{6}$$

Finally, the overall portfolio weight of individual risky asset n is given by

$$\begin{aligned} w_{i,t}^{n} =& w_{i,t}^{*} \cdot w_{i,t}^{n|*} \\ =& \frac{\delta_{i,t}^{*}}{1+\delta_{i,t}^{*}} \cdot \frac{\delta_{i,t}^{n}}{\sum_{n \in N_{i,t}} \delta_{i,t}^{n}} \end{aligned} \tag{7}$$

which depend on the aggregate decision between risk-free bonds and risky assets and the preference across individual risky assets.

2.3 Short Sellers and Market Clearing

To complete the demand system, I model a separate representative short seller, partly due to the tractability of the model and the limited availability of short interest data. Additionally, short sellers are mainly specialized hedge funds with a primary focus on short sales, and they are unlikely to be households, banks, insurances, or mutual funds. As An et al. (2021) point out, these investors shy away from short sales for various reasons, which include but are not limited to regulatory constraints, client restrictions, lack of short-selling talent and experience, large marginal costs and risks of short selling, volatile flows and flow-induced

liquidation costs, and low benchmark-adjusted returns.

Specifically, the representative short seller is modelled similarly to the long investors. The major difference is that, instead of allocating their wealth between the risk-free bond and risky assets, the short seller has zero total wealth. Consequently, they short sell risky assets and invest all proceedings in the risk-free bond. Given total short positions taken on risky assets, the conditional portfolio weights still follow the characteristics-based demand (Equations 2 and 3). However, the capital invested in the risk-free bond and risky assets satisfies $A^0_{-1,t} = -A^*_{-1,t}$, which is modelled as

$$A^*_{-1,t} = -\exp\left(\alpha_{0,-1,t} + \alpha_{1,-1,t} SR_t + \sum_{k\geq 2} \alpha_{k,-1,t} y^*_{k,t} + \theta_{-1} \Gamma^*_{-1,t} + \tilde{\xi}^*_{-1,t}\right) \tag{8}$$

where $\Gamma^*_{-1,t} = \log(\sum_{n\in N_{-1,t}} \delta^n_{-1,t})$ is the diversification value of risky assets for short sellers, parallel to that of a long investor. Equation 8 specifies that, to accommodate zero total wealth of short sellers, the total capital from short sales (rather than the relative weight of long positions on risky assets) depends on the aggregate market conditions and the diversification value.

With the short sellers, I can complete the two-tier asset demand system. Let the bold letter $\mathbf{p}_t$ be the $(N-1)\times 1$ stacked log price vector of individual risky assets excluding the base risky asset $n = 1$. Then, the market clearing conditions for individual risky asset $n = 1, ..., N$ are

$$ME^n_t = \sum_{i=0}^{I} A_{i,t} \cdot w^*_{i,t}(SR_t, \mathbf{p}_t) \cdot w^{n|*}_{i,t}(\mathbf{p}_t) + A^*_{-1,t}(SR_t, \mathbf{p}_t) \cdot w^{n|*}_{-1,t}(\mathbf{p}_t) \tag{9}$$

where I define $w^{n|*}_{i,t} = 0$ when asset n is not in investor i's investment universe $N_{i,t}$ for notational convenience. By the budget constraints, market clearing for the risk-free bond and the aggregate risky asset market are automatically satisfied

$$ME^*_t = \sum_{i=0}^{I} A_{i,t} \cdot w^*_{i,t}(SR_t, \mathbf{p}_t) + A^*_{-1,t}(SR_t, \mathbf{p}_t) \tag{10}$$

Essentially, the market clearing states that the market equity of each asset must equal the total capital invested in said asset across all investors' portfolios. Equation 9 demonstrates

that a shock to individual asset demand could impact asset prices through two channels. The first channel can be viewed as a substitution effect. When the demand shock changes the relative taste among risky assets, the investor tilts their conditional portfolio within the risky asset class towards more desirable assets through $w_{i,t}^{n|*}$. The second channel can be viewed as a wealth effect. As the demand shock affects the diversification value of the risky asset class, the investor redistributes the total capital allocated to the risky asset class through $w_{i,t}^{*}$.

Note that market clearing 9 is a system of N equations in aggregate Sharpe ratio and N asset prices. In Section 5, I solve market clearing conditional on aggregate Sharpe ratio so that the system is exactly determined.

2.4 Demand Elasticities

The nested logit demand system provides a straightforward way to compute demand elasticities, allowing for a more flexible substitution pattern compared to the characteristics-based demand system of Koijen and Yogo (2019).

For long investor $i \geq 0$, let $a_{i,t}^{*} = \log(A_{i,t} w_{i,t}^{*})$ be the log demand of the risky asset class implied by the two-tier demand system, and let $\mathbf{a}_{i,t} = \log(A_{i,t} w_{i,t}^{*} \mathbf{w}_{i,t}^{\cdot|*})$ be that of individual assets with strictly positive holdings in their investment universe excluding the base risky asset $n = 1$. Similarly, define log short position $a_{-1,t} = \log(|A_{-1,t}^{*}|)$ and $\mathbf{a}_{-1,t} = \log(|A_{-1,t}^{*}| \mathbf{w}_{-1,t}^{\cdot|*})$ for the short seller. Then, the demand elasticities with respect to individual asset prices are

$$
\begin{aligned}
\frac{\partial a_{i,t}^{*}}{\partial \mathbf{p}_t'} &= \theta \beta_{1,i,t} (1 - w_{i,t}^{*}) \mathbf{w}_{i,t}^{\cdot|*} \\
\frac{\partial a_{-1,t}^{*}}{\partial \mathbf{p}_t'} &= \theta_{-1} \beta_{1,-1,t} \mathbf{w}_{-1,t}^{\cdot|*} \qquad (11) \\
\frac{\partial \mathbf{a}_{i,t}}{\partial \mathbf{p}_t'} &= \beta_{1,i,t} \operatorname{diag}(\mathbf{w}_{i,t}^{\cdot|*})^{-1} \left((\theta(1 - w_{i,t}^{*}) - 1) \mathbf{w}_{i,t}^{\cdot|*} \mathbf{w}_{i,t}^{\cdot|*\prime} + \operatorname{diag}(\mathbf{w}_{i,t}^{\cdot|*}) \right) \\
\frac{\partial \mathbf{a}_{-1,t}}{\partial \mathbf{p}_t'} &= \beta_{1,-1,t} \operatorname{diag}(\mathbf{w}_{-1,t}^{\cdot|*})^{-1} \left((\theta_{-1} - 1) \mathbf{w}_{-1,t}^{\cdot|*} \mathbf{w}_{-1,t}^{\cdot|*\prime} + \operatorname{diag}(\mathbf{w}_{-1,t}^{\cdot|*}) \right) \qquad (12)
\end{aligned}
$$

which depend on preference over market equity $\beta_{1,i,t}$, preference over diversification value θ (which also governs the substitution pattern among asset classes) and current portfolio weights.

The demand elasticities with respect to aggregate Sharpe ratio are

$$\begin{aligned}\frac{\partial a^*_{i,t}}{\partial SR_t} =& \frac{\partial a^n_{i,t}}{\partial SR_t} = \alpha_{1,i,t}(1 - w^*_{i,t}) \\ \frac{\partial a^*_{-1,t}}{\partial SR_t} =& \frac{\partial a^n_{-1,t}}{\partial SR_t} = \alpha_{1,-1,t}\end{aligned} \tag{13}$$

which depend on investor preference $\alpha_{1,i,t}$ and the current portfolio weight of the risky asset class.

Clearly, the demand elasticities vary across different investors, therefore accommodating heterogeneous substitution patterns. For the aggregate log demand $a^*_t = \log(\sum_{i=-1}^{I} A^*_{i,t})$ and $a^n_t = \log(\sum_{i=-1}^{I} A^n_{i,t})$, the corresponding elasticities are simply the holdings-weighted average of investor-specific elasticities.

3 Data

3.1 Asset Characteristics

Form 13F requires that institutional investment managers report holdings on "section 13F securities", which are largely comprised of equity securities that trade on an exchange (e.g., NYSE, AMEX, NASDAQ). Other reportable assets include certain equity options and warrants, shares of closed-end investment companies, shares of ETFs, and certain convertible debt securities. Because the equity securities of the CRSP-Compustat universe account for approximately 75% of 13F securities in the sample from 1980 to 2019, I focus my analysis on stocks and define the remaining assets and stocks with missing characteristics as the base asset 1 in the risky asset class.

The data on stock prices, dividends, returns, and shares outstanding are from the Center for Research in Security Prices (CRSP) Monthly Stock Database. In the case of missing CRSP data, the Thomson Reuters Institutional Holdings Database (s34 file) is used as an alternative source of data on stock prices and shares outstanding. The two databases generally agree among their shared coverage, but if not, I prioritize the CRSP data. Accounting data are from the Compustat North America Fundamentals Annual and Quarterly Databases. Following standard practice, I merge the two datasets according to the CRSP/Compustat Merged (CCM) link table and ensure that the accounting data were public on the trading date.

Building on the literature (Fama and French (1992); Hou et al. (2015); Koijen and

Yogo (2019)), asset characteristics x_t in my specification include log market equity, log book equity, profitability, investment, dividends to book equity, and market beta. This set of characteristics is highly relevant in the context of asset demand, generates sufficient explanatory power in the cross sections of returns, and does not prompt major issues of overfitting.

I construct the characteristics following Fama and French (2015), which I briefly summarize here. A more detailed description can be found in Appendix A. Profitability is computed as operating profits over book equity. Investment is the annual log growth rate of assets. Dividends to book equity is the ratio of annual dividends per split-adjusted share times shares outstanding to book equity. Market beta is estimated via a 60-month rolling window regression of monthly excess return onto market excess return, where the data on risk-free rate and market excess return are from the Kenneth R. French Data Library. For each time period, I winsorize profitability, investment, and market beta at 2.5% and 97.5% to remove extreme outliers. Similarly, the non-negative dividends to book equity ratio is winsorized at 97.5%.

3.2 Aggregate Conditions

Aggregate-level data are from basic economics databases including the U.S. department of treasury, Federal Reserve Economic Data (FRED), and Kenneth R. French Data Library. In my specification, I only include the market Sharpe ratio (and a constant) in the set of aggregate market conditions, as it is the most pertinent and impactful (Campbell (2017); Gabaix and Koijen (2021)). I leave other market condition variables for future research to avoid an overly complex model and focus on the effect of Sharpe ratio.

I construct the aggregate Sharpe ratio following Whitelaw (1994) and Tang and Whitelaw (2011), with a GARCH(1,1) estimation on monthly market excess return from 1953 to 2019. The explanatory variables in the mean equation include Baa-Aaa spread, the dividend yield, the one-year Treasury yield and lagged market excess return. The explanatory variables in the variance equation include the one-year Treasury yield and the commercial paper-Treasury spread. I also compute the market Sharpe ratio directly using a 30-month rolling window. The two measures provide similar estimates, where the conditional Sharpe ratio has an average of 18.8% with a standard deviation of 18.1% and the rolling Sharpe ratio has an average of 18.4% with a standard deviation of 21.3%. The ex-post unconditional Sharpe ratio in the sample is 14.4%, which is lower than the average conditional Sharpe ratio due to Jensen's inequality. Note that these Sharpe ratios are of monthly frequency and

the corresponding quarterly Sharpe ratio can be obtained by multiplying $\sqrt{3}$.

3.3 Institutional Holdings

The data on institutional holdings are from the Thomson Reuters Institutional Holdings Database (s34 file), which are compiled from the quarterly filings of SEC Form 13F. The full sample covers from 1980q1 to 2019q4. Another possible source is the FactSet 13F Dataset, which I do not use due to its limited time coverage. Form 13F requires investment managers whose total assets under management exceed $100 million in market value to disclose long positions on reportable securities. Therefore, I do not have data on institution-level short positions. Instead, short interest data are from Compustat North America Supplemental Short Interest File, which covers stock-level aggregate short positions.

I group institutions into six types based on the SEC reports: banks, insurance companies, investment advisors (including hedge funds), mutual funds, pension funds, and other 13F institutions (e.g., endowments, foundations, and nonfinancial corporations). Besides the 13F institutions and short sellers, I define the household sector as the aggregate investor who holds the residual shares between total shares outstanding and the sum of shares held by other investors. I also include in the household sector any institution with less than $10 million of total investment in risky assets, no base risky asset, or no risky assets other than the base asset in the investment universe. Therefore, the household sector represents direct household holdings and small institutional investors.

I merge the institutional holdings data with the asset characteristics data by CUSIP number. I compute each investor's total investment in risky assets as the total market value of risky assets held by them. The conditional portfolio weights within the risky asset class is the market value of individual asset holdings over total investment in risky assets. An investor's active share is computed as the sum of the absolute deviation of conditional portfolio weights from the market-weighted portfolio within their investment universe, which is then divided by 2, as in Gabaix and Koijen (2021). Finally, following Koijen and Yogo (2019), I define the investment universe for each investor as all risky assets that they currently hold or have ever held in the previous 11 quarters. As Koijen and Yogo (2019) note, for the median institution, 94% of assets that are currently held were also held in the previous 11 quarters and going further back does not substantially increase the percentage. In Appendix A, I provide a detailed summary of the 13F institutions in the sample.

4 Estimating the Two-Tier Demand System

The literature on nested logit models has theoretically examined a two-step estimation procedure (Domencich and McFadden (1975); Amemiya (1978); McFadden (1981)) and implemented the procedure extensively in practice (Dubin (1986); Falaris (1987); Forinash and Koppelman (1993)). Building on that, I propose a two-step identification strategy to consistently estimate the two-tier asset demand system. In the first step, I estimate the second-tier demand within the risky asset class (Equation 2) based on conditional portfolio weights and an improved instrument from Koijen and Yogo (2019). In the second step, inspired by Gabaix and Koijen (2021), I log-linearize the first-tier demand between asset classes (Equation 4) to circumvent the obstacle posed by unavailable risk-free holdings data. Then, I utilize the estimated asset-specific latent demand from the first step to construct valid instruments for the aggregate panel estimation. This two-step procedure allows me to exploit both cross-sectional and time-series variation in a dynamic setting.

4.1 Demand within the Risky Asset Class

For the second tier of the demand system, I estimate the demand within the risky asset class through the following equation

$$\frac{w_{i,t}^{n|*}}{w_{i,t}^{1|*}} = \exp\left(\beta_{0,i,t} + \beta_{1,i,t} me_t^n + \sum_{k\geq 2} \beta_{k,i,t} x_{k,t}^n\right)\xi_{i,t}^n \qquad n \in N_{i,t} - \{1\} \tag{14}$$

where the asset-specific latent demand $\xi_{i,t}^n$ could be of value zero to accommodate zero holdings and is normalized to have mean 1 so that $\beta_{0,i,t}$ is identified. Because I retain from literature the assumption that characteristics other than price are exogenous, Equation 14 can thus be interpreted as a non-linear regression model with all exogenous explanatory variables but one, i.e., log market equity me_t^n.

To deal with the endogeneity issue, Koijen and Yogo (2019) propose the following instrument

$$\widehat{me}_{i,t}^n = \log\left(\sum_{j\neq i} A_{j,t}^* \frac{1\{n \in N_{j,t}\}}{\sum_{m=1}^N 1\{m \in N_{j,t}\}}\right) \tag{15}$$

where $A_{j,t}^*$ is investor j's total investment in risky assets, and $1\{m \in N_{j,t}\}$ indicates

whether asset m is inside investor j's investment universe at time t. This measure solely depends on the investment universe of other investors and the distribution of total risky capital, both of which are exogenous under their identifying assumptions. The arguments for the exogeneity of investment universe still hold true in this paper. In practice, investment mandates usually dictate investors' investment universes. These are predetermined rules on assets that fall into investment managers' choice set, may they be regulation constraints, client restrictions, investment styles, or index tracking restraints. Therefore, I maintain the assumption of exogenous investment universe.

However, total investment in risky assets for the current period $A^*_{i,t}$, which Koijen and Yogo (2019) refer to as total wealth, is endogenized in the two-tier demand system. Latent demand shocks affect diversification value, which in turn has an impact on the total capital invested in the aggregate risky asset market. For this exact reason, I employ an improved version of the above instrument, substituting current investment in risky assets with average past investment in risky assets

$$\widetilde{me}^n_{i,t} = \log\left(\sum_{j\neq i} \tilde{A}^*_{j,t} \frac{1\{n \in N_{j,t}\}}{\sum_{m=1}^{N} 1\{m \in N_{j,t}\}}\right) \tag{16}$$

where $\tilde{A}^*_{j,t} = \frac{1}{K}\sum_{k=1}^{K} A^*_{j,t-k}$ is the average past investment in risky assets. I choose $K = 3$ when estimating, because a small K might retain a strong correlation with current investment in risky assets, while a large K could diminish the relevance of the instrument. I argue that Equation 16 is a valid instrument for market equity. First, asset-specific latent demand $\xi^n_{i,t}$ only affects the portfolio weight of the risky asset class in current period t but not in previous periods. Using past values reduces the endogenous correlation. Second, I compute the moving average of past investments in risky assets to extract the low-frequency component that is plausibly exogenous to asset-specific demand shock at current time, especially considering that Gabaix and Koijen (2021) observe institutions often have quite stable portfolio weights on risky stocks. Hence, Equation 16 can be viewed as the counterfactual market equity if other investors hold a stable weight of risky assets and allocate risky investments equally within their investment universe. It exploits the fact that an asset accessible to more and larger investors has a larger exogenous component of demand.

With the improved instrument, the identifying assumptions for Equation 14 become

$$\mathbb{E}[\xi_{i,t}^n | \widetilde{me}_{i,t}^n, \tilde{x}_t^n] = 1 \tag{17}$$

where $\tilde{x}_t^n$ includes characteristics other than log market equity.

Finally, with regard to some implementation issues, I follow the suggestions of Koijen and Yogo (2019). (i) I refine the instrument to be more robust by excluding households and short sellers and aggregate only over institutions with little variation in the investment universe, for which at least 95% of assets that are currently held were also held in the previous 11 quarters. (ii) I pool institutions with fewer than 1,000 holdings with similar institutions based on type and total investment in risky assets to estimate their coefficients. (iii) I restrict the coefficient on log market equity of long investors to be less than 1 to guarantee a unique equilibrium defined by market clearing.

4.2 Demand between Asset Classes

For the demand between asset classes of long investors ($i \geq 0$), I start from the following equation inferred from the model (see Equation 4)

$$\frac{P_t^* Q_{i,t}^*}{A_{i,t}^0} = \frac{w_{i,t}^*}{w_{i,t}^0} = \exp\left(\alpha_{0,i,t} + \alpha_{1,i,t} SR_t + \theta \Gamma_{i,t}^* + \tilde{\xi}_{i,t}^*\right) \tag{18}$$

where P_t^* is the aggregate price of the risky asset class, $Q_{i,t}^*$ is the effective number of shares of the aggregate risky asset market held by investor i, and $A_{i,t}^0$ is the risk-free bond holdings. An immediate challenge I encounter when estimating Equation 18 is the lack of access to institutional risk-free holdings. That is, I do not have data on $A_{i,t}^0$. Unlike the regulations on public equity holdings which require comprehensive reporting by 13F institutions, regulations on public bond holdings typically do not require disclosure of holdings or trades. The exceptions to this rule are the bond holdings of insurance companies, which must be disclosed to the National Association of Insurance Commissioners in Schedule D filings, and the bond holdings of selected investment managers (certain mutual funds and pension funds), which must be disclosed to the SEC.

These data, albeit valuable for other studies, do not help with the estimation in the context of an asset demand system including all investors. Therefore inspired by Gabaix and Koijen (2021), instead of directly estimating Equation 18, I apply a log-linearization

which allows me to bypass the need for risk-free holdings data and estimate the parameters of interest only using data on the aggregate risky asset market.

To log-linearize the actual demand equation, consider a simpler "baseline" economy, which is on a balanced growth path. Let $\bar{P}_t^*,\ \bar{Q}_{i,t}^*, \bar{A}_{i,t}^0, \bar{A}_{i,t}$ be the baseline aggregate price, number of shares of the aggregate risky asset market held by investor i, risk-free bond holdings, and total wealth of investor i. Then, with no demand shocks, the baseline portfolio weights of investor i can be expressed as

$$\frac{\bar{P}_t^* \bar{Q}_{i,t}^*}{\bar{A}_{i,t}^0} = \exp\left(\alpha_{0,i,t} + \alpha_{1,i,t}\overline{SR} + \theta\bar{\Gamma}_i^*\right) \tag{19}$$

where $\overline{SR}$ and $\bar{\Gamma}_i^*$ are respectively the baseline aggregate Sharpe ratio and diversification value of the risky asset class, which would be constant in a balanced economy and can be interpreted as the corresponding long-term averages in the real economy.

From the baseline portfolio weights, two types of shocks occur, and investors shift their holdings accordingly to realize the actual weights. One is demand shocks, which deviate asset prices from baseline, and the other is an unexpected inflow shock to the investor's total wealth $\bar{A}_{i,t}$. That is, before the shocks, investor i allocates their wealth between the asset classes with baseline weights and satisfies the budget constraint $\bar{A}_{i,t} = \bar{P}_t^* \bar{Q}_{i,t}^* + \bar{A}_{i,t}^0$. Then shocks strike, and, before the investor makes any adjustment to their holdings, the budget becomes $A_{i,t} = P_t^* \bar{Q}_{i,t}^* + \bar{A}_{i,t}^0 + F_{i,t}$, where $F_{i,t}$ is an unexpected inflow into investor i's total wealth. Faced with deviated prices and total wealth, investor i eventually adjusts their holdings following Equation 18 with the actual budget constraint $A_{i,t} = P_t^* Q_{i,t}^* + A_{i,t}^0$.

Therefore, dividing Equation 18 by Equation 19, the demand between asset classes can be expressed in terms of deviation from baseline

$$(1 + \Delta p_t^*)(1 + \Delta q_{i,t}^*) = (1 + \Delta a_{i,t}^0) \exp\left(\tilde{\alpha}_{0,i,t} + \alpha_{1,i,t} SR_t + \theta\Gamma_{i,t}^* + \tilde{\xi}_{i,t}^*\right)$$

where $\Delta p_t^* = (P_t^* - \bar{P}_t^*)/\bar{P}_t^*,\ \Delta q_{i,t}^* = (Q_{i,t}^* - \bar{Q}_{i,t}^*)/\bar{Q}_{i,t}^*,\ \Delta a_{i,t}^0 = (A_{i,t}^0 - \bar{A}_{i,t}^0)/\bar{A}_{i,t}^0$ are the percentage deviation from baseline, and $\tilde{\alpha}_{0,i,t} = -\alpha_{1,i,t}\overline{SR} - \theta\bar{\Gamma}_i^*$. Plugging in the budget constraints $A_{i,t} = P_t^* \bar{Q}_{i,t}^* + \bar{A}_{i,t}^0 + F_{i,t}$ and $A_{i,t} = P_t^* Q_{i,t}^* + A_{i,t}^0$ and log-linearizing around the baseline economy, investor i's demand change follows

$$\Delta p_t^* + M_{i,t}\Delta q_{i,t}^* = \tilde{\alpha}_{0,i,t} + \alpha_{1,i,t} SR_t + \theta \Gamma_{i,t}^* + \bar{\xi}_{i,t}^* \tag{20}$$

where $M_{i,t} = 1 + \exp(\alpha_{0,i,t} + \alpha_{1,i,t}\overline{SR} + \theta\bar{\Gamma}_i^*)$ partially corresponds to the price multiplier in Gabaix and Koijen (2021) whose reciprocal measures the elasticity in the aggregate risky asset market, and $\bar{\xi}_{i,t}^* = \tilde{\xi}_{i,t}^* + F_{i,t}/\bar{A}_{i,t}^0$ is the approximate aggregate latent demand. Because Equation 20 is expressed in terms of risky assets holdings rather than risk-free bond holdings, the equation is estimable empirically with valid instruments. In practical implementation, I compute the percentage deviation Δp_t^* and $\Delta q_{i,t}^*$ as the percentage deviation from the previous period following Gabaix and Koijen (2020), and winsorize $\Delta q_{i,t}^*$ at 1% and 99% to minimize the impact of extreme outliers. I also parameterize the coefficients $M_{i,t}$, $\tilde{\alpha}_{0,i,t}$ and $\alpha_{1,i,t}$ as functions of a linear combination of the investor's characteristics. Per suggestion of Gabaix and Koijen (2021), an investor's characteristics include average past active share, which is a measure of the activeness of management; the log of average past investment in risky assets, which is a measure of size; and the investor's type.

Now, the lone obstacle to estimation is the endogeneity of prices. In the two-tier demand system, the aggregate latent demand $\tilde{\xi}_{i,t}^*$ partially determines both asset prices and aggregate Sharpe ratio in the equilibrium. Therefore, I need instruments for both of these variables, as well as the diversification value $\Gamma_{i,t}^*$ which is also a function of individual asset prices. Exploiting the fact that asset-specific latent demand $\xi_{i,t}^n$ is, by assumption of the model, orthogonal to the aggregate latent demand $\tilde{\xi}_{i,t}^*$, I can therefore construct valid instruments based on $\xi_{i,t}^n$ from the first step of estimation. I propose the following two instruments for estimating Equation 20

$$\tilde{z}_{i,t} = \sum_{n \in N_{i,t} - \{1\}} w_{m,i,t}^n \log \xi_{i,t}^n \tag{21}$$

$$\tilde{z}_t = \sum_{i=1}^{I} S_{i,t-1} \tilde{z}_{i,t} \tag{22}$$

where $w_{m,i,t}^n = \frac{ME_{t-1}^n}{\sum_{n \in N_{i,t}-\{1\}} ME_{t-1}^n}$ is the market weights of risky assets inside investor i's investment universe at time t, and $S_{i,t-1} = \frac{A_{i,t-1}^*}{\sum_i A_{i,t-1}^*}$ is the market share of risky assets held by investor i at time $t-1$.

Theoretically, other functions of asset-specific latent demand could also be valid instru-

ments because of the imposed orthogonality. I specifically employ the weighting scheme in Equations 21 and 22 to enhance the relevance of the instruments and address possible weak instrument issues. The idea is similar to the granular instrumental variable of Gabaix and Koijen (2020). Asset-specific demand shock of investors with larger market share and of assets with larger market equity would have a larger exogenous impact on equilibrium prices. Therefore, the above weighting scheme allows me to extract an adequately relevant component in asset prices that is also exogenous to aggregate demand shock. Analogous to the construction of $\widetilde{me}^n_{i,t}$ in 16, I refine the instrument $\tilde{z}_t$ to be more robust by excluding households and short sellers.

As for the diversification value $\Gamma^*_{i,t}$ which depends on the risky asset investment diversified over market equity, other characteristics and asset-specific latent demand, the only endogenous component is the market equity. Therefore, I construct a counterfactual diversification value that only depends on diversification over characteristics excluding market equity and asset-specific latent demand

$$\tilde{\Gamma}^*_{i,t} = \log\left(1 + \sum_{n \in N_{i,t} - \{1\}} \exp\left(\beta_{0,i,t} + \sum_{k \geq 2} \beta_{k,i,t} x^n_{k,t}\right) \xi^n_{i,t}\right) \tag{23}$$

Note that the sole difference between $\Gamma^*_{i,t}$ and $\tilde{\Gamma}^*_{i,t}$ is that the effect of market equity is excluded in the proposed instrument $\tilde{\Gamma}^*_{i,t}$ to ensure exogeneity.

Finally, with all the proposed instruments, I can estimate Equation 20 using the panel of investors with the following identifying assumptions

$$\mathbb{E}[\bar{\xi}^*_{i,t} | \tilde{z}_t, \tilde{z}_{i,t}, \tilde{\Gamma}^*_{i,t}] = 0 \tag{24}$$

With regard to the short seller, the demand can be directly estimated by Equation 8 using the instruments $\tilde{z}_t$ and $\tilde{\Gamma}^*_{-1,t}$ proposed above. Notably, as a result of the aggregation of short positions, total short interest has a clear time trend. To accommodate this observation, a time variable is added to the constant $\alpha_{0,-1,t}$ in Equation 8 in addition to the investor's characteristics. This brings the identifying assumptions to

$$\mathbb{E}[\tilde{\xi}^*_{-1,t} | \tilde{z}_t, \tilde{\Gamma}^*_{-1,t}] = 0 \tag{25}$$

To examine the performance of the above proposed instruments and estimation procedure, I conduct a simulation study in a simplified setting in Appendix B. The results show that the procedure reliably uncovers all aggregate coefficients in an accurate manner.

4.3 Estimated Two-Tier Demand System

Figure 1 summarizes the coefficients on asset characteristics in the demand within the risky asset class (Equation 14) from 1981q1 to 2019q4. For each investor type, I report the cross-sectionally averaged coefficients, weighted by total investment in risky assets. To ease interpretation, Figure 1 reports the coefficients on log market-to-book ratio $\beta_{1,i,t}$ and log book equity $\beta_{1,i,t} + \beta_{2,i,t}$ instead of $\beta_{1,i,t}$ and $\beta_{2,i,t}$. Overall, the estimates are reasonably consistent with Koijen and Yogo (2019) with several notable differences.

Equation 12 implies that the long investor has a less elastic demand as the coefficient on log market-to-book ratio approaches 1. Therefore, Figure 1 shows that the demand of mutual funds is less elastic compared to other institutions for most of the sample periods, while that of investment advisors is among the most elastic. Over time, banks, insurance companies and pension funds are subject to less elastic demand, contrary to households. This could result from the fact that financial intermediaries have been growing in size and their asset management has become more passive and subject to benchmarking in order to attract a wider range of clientele. For short sellers, the coefficient on log market-to-book ratio is mostly above 1, highlighting their role in detecting over-valued assets.

The coefficient on log book equity captures the demand for size. Banks, insurance companies and mutual funds have a growing fondness for stocks with larger size, in contrast with investment advisors and households. This is consistent with the findings of Blume and Keim (2012) and Koijen and Yogo (2019) that larger institutions tend to hold larger stocks in their portfolios. Compared to long investors, short sellers prefer large firms, as they have concluded their high-growth phase.

On average, financial institutions tilt their portfolio more towards stocks with higher profitability, higher investment, lower dividend to book equity ratio and lower market beta than households. This reflects the view that intermediaries interpret information differently and invest in a more calculated way, partially explaining why they have become increasingly popular. Short sellers, as expected, prefer stocks with low profitability, low investment, low dividend to book equity ratio and high market beta. And the coefficients on profitability, dividend to book equity ratio and market beta tend to move in the opposite direction of long investors, suggesting that demand on the long leg and on the short leg are interlinked.

Interestingly, short sellers were overly reactive towards dividends in the first half of the sample compared to the second half, which could be evidence for a transition in short-selling criteria from a singular criterion based on cash flow to a more diverse taste. Also notably, investors tend to hold short positions instead of long positions on stocks with higher market beta during recessions, implying that the demand for market risk is procyclical.

Besides characteristics, an integral part of demand is the latent demand. Figure 2 reports the cross-sectional standard deviation of log asset-specific latent demand by investor type, weighted by total investment in risky assets. A higher standard deviation implies more extreme portfolio weights that are tilted away from observed characteristics. Among institutions, mutual funds have the most extreme portfolios tilted away from observed characteristics, followed closely by insurance companies, investment advisors and banks, while pension funds have the least variation in latent demand. Households, for most of the sample periods, tend to condition their demand more intensively on observed characteristics than institutions. This indicates that intermediaries might have a larger set of information than observed fundamentals, consistent with the view that some institutions are "smart money" investors. Finally, short sellers have quite volatile latent demand, which might be a result of the aggregation of short positions.

As for demand between asset classes, Table 1 and Figure 3 summarize the estimated coefficients in Equation 20 and the estimated equity weights (i.e., portfolio weights on risky assets that are mainly equities). Table 1 clearly shows that investors value the diversification of their portfolios. The coefficient on the diversification value θ is estimated to be 0.06 for long investors and 0.31 for short sellers, both of which are significant.

In Figure 3, the cross-sectionally averaged coefficients and equity weights are reported for each investor type, weighted by total investment in risky assets. Mutual funds are the most reactive towards the aggregate Sharpe ratio, followed by banks and pension funds. Households, insurance companies and investment advisors are less likely to tilt their portfolios away from equities based on the market condition. Short sellers, as expected, conduct fewer short sales when the aggregate stock market is in a good state and equities are expected to perform well. Over time, short sellers have become more reactive to aggregate Sharpe ratio.

The price multiplier partially measures the demand elasticity in the aggregate market when holding the Sharpe ratio constant. The average price multiplier across investors and time is 4.3, which is slightly smaller than Gabaix and Koijen (2021)'s estimate of 5. This lends further evidence that the aggregate stock market is surprisingly inelastic. However, different types of investors contribute to the inelastic market to various degrees. Insurance

companies and mutual funds are among the least elastic, possibly due to their preference for equities and size-related benchmarking. The other institutions have more elastic demand, while households are remarkably elastic. It is possible that households, subject to fewer restrictions and investment objectives, have less transaction cost when redistributing their wealth among different asset classes.

Consistent with the observation made by Gabaix and Koijen (2021), investors hold a quite stable equity share out of their total wealth over time. Note that the estimated equity share only covers equity securities that are directly held by investors and are reportable on Form 13F. The actual portfolio weight on equities could be higher than depicted in Figure 3. This is especially true for households because only direct holdings by households are used in the estimation, but they could hold equities indirectly through intermediaries. Despite this concern, the estimated equity share matches well with the available data on mutual funds and pension funds. On average, mutual funds and pension funds are estimated to invest 80% and 67%, respectively, of their wealth in equities. In comparison, data show that mutual funds hold approximately 82–83% of equities on average, while pension funds hold approximately 67–68% of equities.

Finally, Table 2 reports the standard deviation and auto-correlation of the estimated aggregate latent demand by investor type. Compared to asset-specific latent demand, aggregate latent demand has a smaller variation, and the standard deviation is fairly consistent across different types of long investors. It is also highly auto-correlated, especially for institutions, partly because the aggregate latent demand carries some persistent aggregate factors or unobserved investor expectations on the stock market. These facts could shed some light on why institutions hold such a stable equity share.

5 Applications

In the two-tier asset demand system, log prices are entirely determined through market clearing 9 by an implicit function

$$\mathbf{p}_t = g(\mathbf{q}_t, \tilde{\mathbf{x}}_t, \mathbf{A}_t, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) \tag{26}$$

where $\mathbf{q}_t$ is an $(N-1)\times 1$ vector of log number of shares outstanding for individual assets, $\tilde{\mathbf{x}}_t$ is an $(N-1)\times(K_x-1)$ matrix of asset characteristics other than log market equity, $\mathbf{A}_t$ is an $(I+1)\times 1$ vector of long investors' total wealth, α_t is an $(I+2)\times 3$ matrix of

aggregate coefficients and aggregate Sharpe ratio, β_t is an $(I+2) \times K_x$ matrix of coefficients on asset characteristics, $\tilde{\xi}_t^*$ is an $(I+2) \times 1$ vector of aggregate latent demand, and ξ_t is an $(N-1) \times (I+2)$ matrix of asset-specific latent demand.

Equation 26 is repeatedly used in the following four asset pricing applications. First, I decompose return variance in the stock market into supply- and demand- side effects. Second, I inspect the connection between institutional demand and common return premiums. Third, I examine the asset pricing role of short sales by studying the effect of short sales on return volatility and stock valuation. Finally, I use the model to investigate the underlying channel of the return predictability of dividend-price ratio in the aggregate stock market.

5.1 Variance Decomposition of Stock Returns

Literature has long asked what contributes to the observed variation in stock returns. The findings of Fama and MacBeth (1973) suggest stock characteristics are an important source of variation, while Gompers and Metrick (2001) and Koijen and Yogo (2019) point out that institutional demand is a key factor in explaining return volatility. Following Koijen and Yogo (2019), I decompose return variance into supply- and demand-side effects using the two-tier demand system and attempt to offer some insight. I start with the vector of log stock returns, defined as

$$\mathbf{r}_{t+1} = \mathbf{p}_{t+1} - \mathbf{p}_t + \mathbf{v}_{t+1}$$

where $\mathbf{v}_{t+1} = \log(1 + \exp(\mathbf{d}_{t+1} - \mathbf{p}_{t+1}))$ and $\mathbf{d}_{t+1}$ is the vector of log dividends per share at time $t+1$. The capital gain can be decomposed into

$$\mathbf{p}_{t+1} - \mathbf{p}_t = \Delta\mathbf{p}(\mathbf{q}) + \Delta\mathbf{p}(\mathbf{x}) + \Delta\mathbf{p}(\mathbf{A}) + \Delta\mathbf{p}(\alpha) + \Delta\mathbf{p}(\beta) + \Delta\mathbf{p}(\tilde{\xi}^*) + \Delta\mathbf{p}(\xi)$$

where

$$
\begin{aligned}
\Delta\mathbf{p}(\mathbf{q}) =& g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_t, \mathbf{A}_t, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) - g(\mathbf{q}_t, \tilde{\mathbf{x}}_t, \mathbf{A}_t, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) \\
\Delta\mathbf{p}(\mathbf{x}) =& g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_t, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) - g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_t, \mathbf{A}_t, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) \\
\Delta\mathbf{p}(\mathbf{A}) =& g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) - g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_t, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) \\
\Delta\mathbf{p}(\alpha) =& g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_{t+1}, \beta_t, \tilde{\xi}_t^*, \xi_t) - g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_t, \beta_t, \tilde{\xi}_t^*, \xi_t) \\
\Delta\mathbf{p}(\beta) =& g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_{t+1}, \beta_{t+1}, \tilde{\xi}_t^*, \xi_t) - g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_{t+1}, \beta_t, \tilde{\xi}_t^*, \xi_t) \\
\Delta\mathbf{p}(\tilde{\xi}^*) =& g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_{t+1}, \beta_{t+1}, \tilde{\xi}_{t+1}^*, \xi_t) - g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_{t+1}, \beta_{t+1}, \tilde{\xi}_t^*, \xi_t) \\
\Delta\mathbf{p}(\xi) =& g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_{t+1}, \beta_{t+1}, \tilde{\xi}_{t+1}^*, \xi_{t+1}) - g(\mathbf{q}_{t+1}, \tilde{\mathbf{x}}_{t+1}, \mathbf{A}_{t+1}, \alpha_{t+1}, \beta_{t+1}, \tilde{\xi}_{t+1}^*, \xi_t)
\end{aligned}
$$

All counterfactual prices are computed numerically through market clearing 9. Then, I decompose the cross-sectional variance of log returns as

$$
\begin{aligned}
\mathrm{Var}(\mathbf{r}_{t+1}) =& \mathrm{Cov}(\Delta\mathbf{p}(\mathbf{q}), \mathbf{r}_{t+1}) + \mathrm{Cov}(\Delta\mathbf{p}(\mathbf{x}), \mathbf{r}_{t+1}) + \mathrm{Cov}(\mathbf{v}_{t+1}, \mathbf{r}_{t+1}) \\
&+ \mathrm{Cov}(\Delta\mathbf{p}(\mathbf{A}), \mathbf{r}_{t+1}) + \mathrm{Cov}(\Delta\mathbf{p}(\alpha), \mathbf{r}_{t+1}) + \mathrm{Cov}(\Delta\mathbf{p}(\beta), \mathbf{r}_{t+1}) \\
&+ \mathrm{Cov}(\Delta\mathbf{p}(\tilde{\xi}^*), \mathbf{r}_{t+1}) + \mathrm{Cov}(\Delta\mathbf{p}(\xi), \mathbf{r}_{t+1}) \qquad (27)
\end{aligned}
$$

The first three terms in Equation 27 capture the supply-side effects arising from changes in shares outstanding, stock characteristics and the dividend yield. The next term represents the wealth effect on the demand side due to change in total wealth. The last four terms reflect the demand-side effects stemming from investors' taste, including preference between asset classes and among individual assets.

Table 3 presents the variance decomposition of annual stock returns for the full sample period from 1982 to 2019 and for the latter half of the sample from 2000 to 2019. Annual stock returns are calculated at the end of the second quarter, as many firms update their characteristics in June.

In the full sample, supply-side effects explain 10.7 percent of stock return variation in total, where shares outstanding, characteristics and dividend yield account for 1.0 percent, 9.2 percent and 0.5 percent respectively. On the demand side, the wealth effect due to change in total wealth explains 2.5 percent. Interestingly, aggregate preference only accounts for 0.6 percent of return volatility, most of which is due to changes in aggregate latent demand. This is possibly a result of the low variation and high auto-correlation of aggregate latent demand, which also explains the stable portfolio weight on equity over time. On the other

hand, individual stock preference accounts for a total of 86.2 percent, where only 3.3 percent is due to changes in coefficients. Asset-specific latent demand is unequivocally the most significant source of variation in stock returns. The extensive margin, which captures changes in investment universe, explains 14.0 percent, while the intensive margin, which captures changes in portfolio weights within the investment universe, explains 69.0 percent.

In comparison, during the post-2000 period, the effect of aggregate preference has more than doubled, indicating a more flexible demand for equities. Also of interest is that the wealth effect due to changes in investors' total wealth has increased sharply. This is consistent with the rapidly expanding size of institutions and their growing roles in asset pricing.

The variance decomposition provides further support for the findings of Koijen and Yogo (2019). Observed characteristics have low explanatory power in stock returns. Instead, investor sentiment and disparity in latent preference is the key source for explaining return volatility.

5.2 Intermediaries and Return Premiums

Besides return variation, observed anomaly returns have also been of continuous interest in asset pricing research. Since Fama and French (1992), many characteristics-based premiums have been studied (e.g., Carhart (1997); Novy-Marx (2013); Fama and French (2015)). One important question is which return anomalies can be attributed to intermediaries and which are more fundamentally founded.

I investigate the connection between common characteristics-based premiums and institutional demand by computing counter-factual return spreads when financial intermediaries become neutral towards the corresponding characteristics. That is, I impose the corresponding coefficient on the characteristic in Equation 14 to be zero for all institutions. Following Fama and French (1992) and Fama and French (2015), I sort common stocks traded on NYSE, AMEX, or NASDAQ into deciles based on said characteristics and compute the average return spreads between the top decile and the bottom decile. The returns are computed annually from January to December, in accordance with the methodology of Kenneth R. French Data Library.

Panel A of Table 4 reports the average annual return spreads and the counter-factual values from 1982 to 2019. In the data, I observe notable premiums for all four characteristics: size, value, profitability and investment. In contrast, when financial intermediaries cease to tilt their portfolio based on the corresponding characteristics, size premium and investment premium can no longer be observed, while value premium is reduced by approximately 40%.

Profitability premium is largely unaffected both in terms of magnitude and significance.

These results indicate that size premium can be mainly attributed to intermediaries, which is consistent with the findings of Blume and Keim (2012) that institutions prefer larger firms. For value premium, though institutional demand might be an important factor, fundamental reasons such as low-frequency cash flow growth, growth opportunities, and exposure to innovation displacement risk, cannot be discounted. Profitability premium does not arise from institutional demand. A possible explanation for this anomaly return could be industrial competition proposed by Dou et al. (2021). Finally, intermediaries could account for the majority of investment premium, as they prefer aggressively expanding firms. Overall, these findings could offer some insight into how financial intermediaries contribute to the return anomalies in the stock market. As intermediaries' preference over certain characteristic shifts institutional demand, stock returns adjust accordingly and generate observable return premiums.

On the other hand, while the literature on return anomalies predominantly focus on the long leg, several papers have inspected the short leg of returns and attributed several anomalies, at least partially, to short-sale constraints. For instance, Stambaugh et al. (2012) show that with short-sale impediments, the short leg of various anomalies are more profitable following high investor sentiment. Avramov et al. (2013) find that several anomaly returns are derived from taking short positions in high credit risk firms which they argue may be hard to short sell, while Drechsler and Drechsler (2014) observe that high short-fee stocks predominate in the short leg of anomalies and drive their returns. I examine this claim by computing the counter-factual return spreads when short sales are banned, which are reported in Panel B of Table 4. The results show that tightening the short-sale constraints does not generate larger or more significant return premiums, indicating that the short leg might not be the main driver of observed anomaly returns.

5.3 Role of Short Sales

Economists generally agree on the beneficial role of short sellers for unearthing over-valued stocks and improving the efficiency of the market. However, there exist several debates on the exact role of short sales due to conflicting evidence. One debate is on whether short-sale constraints contributed to the rise of the dot-com bubble. While Ofek and Richardson (2003) argue short-sale restrictions were crucial for the bubble and Haruvy and Noussair (2006) demonstrate loosening the restrictions could induce prices to approach fundamental values, Geczy et al. (2002) and Battalio and Schultz (2006) show that these constraints were

not binding. Another debate is about the effectiveness of a short-sale ban to inflate stock prices, especially during a financial crisis, since the SEC issued a temporary shorting ban at the worst of the 2008 crisis. Boehmer et al. (2013) observe that small-cap stocks were largely unaffected, whereas large-cap stocks subject to the ban witnessed a jump in price. However, Battalio et al. (2012) find the ban failed to support stock prices. Beber and Pagano (2013) study short-sale bans in different countries and find that these bans are not associated with better performance of stock prices, except possibly for the U.S. financial market.

In light of these diverging views on the actual role of short sales, I conduct three experiments to lend some evidence from the perspective of a demand system. First, I decompose return variance into long- and short-side effects, in a similar fashion as the first application. I modify Equation 27 to

$$\begin{aligned}\mathrm{Var}(r_{t+1}) =& \mathrm{Cov}(\Delta\mathbf{p}(\mathbf{q}) + \Delta\mathbf{p}(\mathbf{x}) + v_{t+1}, r_{t+1}) \\ &+ \sum_{i=0}^{I} \mathrm{Cov}(\Delta\mathbf{p}(\mathbf{A_i}) + \Delta\mathbf{p}(\alpha_\mathbf{i}) + \Delta\mathbf{p}(\beta_\mathbf{i}) + \Delta\mathbf{p}(\tilde{\xi}^*_\mathbf{i}) + \Delta\mathbf{p}(\xi_\mathbf{i}), r_{t+1}) \\ &+ \mathrm{Cov}(\Delta\mathbf{p}(\mathbf{A_{-1}}) + \Delta\mathbf{p}(\alpha_\mathbf{-1}) + \Delta\mathbf{p}(\beta_\mathbf{-1}) + \Delta\mathbf{p}(\tilde{\xi}^*_\mathbf{-1}) + \Delta\mathbf{p}(\xi_\mathbf{-1}), r_{t+1})\end{aligned}$$

The first term is the total supply-side effect. The second term is the sum of the demand-side effects due to all long investors. The third term is the short-side effects. Figure 4 reports the proportion of cross-sectional return variance explained by the change in short sellers. In the first half of the sample, short sellers played an insignificant role in explaining variation in stock returns. Of interest is the time when short sellers first significantly accounted for a part of return volatility. It coincides with the dot-com bubble, reflecting the active role short sellers play in uncovering over-valued stocks. In the second half of the sample, the importance of short sales continues to grow and explains over 4 percent of return variation in 2008. But following the short-sale ban in September 2008, the impact of short sales temporarily drops, only to bounce back in 2011. Eventually, it stabilizes at explaining around 3 percent of return variance as the economy enters a boom.

The variance decomposition highlights the growing importance of short sales, especially during financially abnormal times. However, it does not tell us how short sales affect the valuation of stocks. In the second experiment, I investigate the role of short sales and short-sale constraints during the dot-com bubble by computing counter-factual stock indices in a world where there is no shorting or more aggressive shorting. In the first case, I remove short

sellers from the demand system. In the second case, I allow more aggressive short-selling activities by increasing the intercept term in Equation 8, in order to imitate the outcome of a relaxation on short-sale constraints. I consider two scenarios where short sellers are moderately more aggressive or extremely more aggressive, on average increasing total short positions by 50 percent or 5 times.

Figure 5 compares the real index values (S&P 500 and overall) and the counter-factual ones from 1996q1 to 2003q4. Though short sales indeed have an impact on stock prices, it is within normal range and could not account for the formation of the bubble. Even with extremely aggressive short-selling activities, the bubble shape still exists. Figure 6 demonstrates that similar patterns are observed for large stocks and growth stocks, which were the main components of the bubble shape. Therefore, I conclude that short-sale constraints were unlikely to be a crucial factor for the dot-com bubble, supporting the findings of Geczy et al. (2002) and Battalio and Schultz (2006).

In the third experiment, I further examine the price impact of short sales by studying the repricing effect of a short-sale ban on stocks with different characteristics. Specifically, I sort stocks into deciles based on each characteristic and evaluate the average change in valuation for each decile with the following repricing measure revised from Koijen et al. (2019)

$$RP_{-1,j} = \frac{1}{T}\sum_{t}\left(\frac{\sum_{n\in Q_j}(ME_t^{n,CF} - ME_t^n)}{\sum_{n\in Q_j} ME_t^n}\right) \tag{28}$$

where $ME_t^{n,CF}$ is the counter-factual market equity of asset n without short sales, and Q_j denotes the j-th decile. The repricing measures the change in total market equity in each decile when a short-sale ban is in effect.

Tables 5 and 6 report the average repricing around the 2008 financial crisis and in the full sample. During the crisis, the short-sale ban significantly inflates stock prices, especially for large firms, consistent with the observation by Boehmer et al. (2013). However, the repricing effect on ultra large firms is minimal, likely due to a combination of reasons including higher resilience towards shorting and the dominating scale of long investors. On the other hand, the ban undermines the performance of small-cap stocks, implying that short sellers are more correctional in terms of uncovering over-valued stocks than they are predatory. For book-to-market ratio, profitability and investment, total market equity increases for every decile under the short-sale ban. Interestingly, the repricing effect is much more notable for less profitable stocks, indicating that short sellers target distressed low-profitability firms. These

repricing effects are also found in the full sample but on a smaller scale. The asymmetry in repricing on stocks with different characteristics is also reflected in Daniel et al. (2022), though they focus on price momentum as the main characteristic of interest.

5.4 Return Predictability of Dividend-Price Ratio

Campbell and Shiller (1988b) first document the classic pattern that dividend-price (D/P) ratio significantly predicts future returns in the aggregate stock market and increasingly so with longer horizons. Different studies have examined the mechanism behind the return predictability of dividend-price ratio. Campbell and Cochrane (1999) show that the effective risk aversion captured by their consumption surplus utility provides grounds for the connection between dividend-price ratio and expected returns, while Lakonishok et al. (1994) find that irrational forecast on long-run cash flow growth is the essential link. I investigate the connection between the return predictability of dividend-price ratio and aggregate demand by re-inspecting this predictability pattern with simulated asset demand.

In the first simulation, I set $\alpha_{1,i,t} = 0$ in Equation 18 so that investors are neutral towards aggregate Sharpe ratio. This allows me to study how investors' overall preference between risk and return affects the relationship between dividend-price ratio and future returns. In the second simulation, I replace aggregate latent demand $\tilde{\xi}^*_{i,t}$ with idiosyncratic normal shocks of the same volatility to eliminate any effect of unobserved beliefs, including irrational forecast on long-run cash flow growth. Finally, I impose both constraints to examine the combined effects. In each case, I compute the log excess return and log dividend-price ratio of the value-weighted market portfolio. I then inspect the return predictability of dividend-price ratio with regressions of future log excess return on log dividend-price ratio.

Table 7 reports quarterly regression results from 1982q2 to 2019q4 with real and simulated data. The predictability pattern in Campbell and Shiller (1988b) is revealed in the actual data: a high dividend-price ratio is associated with high expected returns, and the scale and significance of the coefficient rise to an impressive level with long horizons. When investors become neutral towards the aggregate Sharpe ratio, the predictability weakens marginally and the magnitude of the coefficients remains unchanged. In comparison, eliminating the effect of aggregate latent demand dilutes the predictability markedly and shrinks the coefficients towards zero. The combined effect is even more prominent, with the second channel accounting for the bulk of it.

This comparison offers some insight into the the mechanism behind the return predictability of dividend-price ratio. Though risk-return balance certainly plays a role, unobserved

preference and subjective beliefs about the aggregate stock market should be recognized as the main driver. Therefore, the theory of irrational forecast on long-run cash flow growth proposed by Lakonishok et al. (1994) provides a more plausible explanation for the tight link between dividend-price ratio and future returns.

6 Conclusion

I develop a two-tier asset demand system that incorporates endogenous aggregate allocation and short sales. The framework nests characteristics-based individual asset demand within the demand for different asset classes and allows for a more flexible substitution pattern across assets. I also propose a two-step estimation procedure with a novel instrument for aggregate estimation that addresses the endogeneity of aggregate price and Sharpe ratio. The dynamic estimation procedure allows me to exploit both cross-sectional and time-series variation in institutional holdings and jointly estimate micro- and macro-elasticities.

The two-tier system could shed light on a broad set of debates related to the role of institutional demand and short sales in both individual and aggregate stock markets. For individual stocks, I find that asset-specific latent demand is the dominating source of return volatility, and institutional demand accounts for a large portion, if not all, of observed return premiums in size, value and investment. Regarding short sales, short-sale constraints cannot explain return anomalies and were not a crucial factor for the formation of the dot-com bubble. However, short sales have significant but disparate pricing impact on stocks with different characteristics. A short-sale ban inflates stock prices, especially for large and less profitable firms but not for small firms. For the aggregate stock market, I find that unobserved aggregate preference and beliefs rather than risk-return balance is the main driver of return predictability of dividend-price ratio. Future work could build on the two-tier demand system to incorporate other asset classes and answer questions about how institutional demand explains movements in aggregate and individual markets jointly.

References

Adrian, T., Etula, E., Muir, T., 2014. Financial intermediaries and the cross-section of asset returns. Journal of Finance 69 (6), 2557–2596.

Aït-Sahalia, Y., Yu, J., 2006. Saddlepoint approximations for continuous-time markov processes. Journal of Econometrics 134 (2), 507–551.

Amemiya, T., 1978. On a two-step estimation of a multivariate logit model. Journal of Econometrics.

An, L., Huang, S., Lou, D., Shi, J., 2021. Why don't most mutual funds short sell? LSE Financial Markets Group.

Asquith, P., Pathak, P. A., Ritter, J. R., 2005. Short interest, institutional ownership, and stock returns. Journal of Financial Economics 78 (2), 243–276.

Avramov, D., Chordia, T., Jostova, G., Philipov, A., 2013. Anomalies and financial distress. Journal of Financial Economics 108 (1), 139–159.

Basak, S., Pavlova, A., 2013. Asset prices and institutional investors. American Economic Review 103 (5), 1728–1758.

Basak, S., Pavlova, A., Shapiro, A., 2007. Optimal asset allocation and risk shifting in money management. Review of Financial Studies 20 (5), 1583–1621.

Battalio, R., Schultz, P., 2006. Options and the bubble. Journal of Finance 61 (5), 2071–2102.

Battalio, R. H., Mehran, H., Schultz, P. H., 2012. Market declines: What is accomplished by banning short-selling? Current Issues in Economics and Finance 18 (5).

Beber, A., Pagano, M., 2013. Short-selling bans around the world: Evidence from the 2007-09 crisis. Journal of Finance 68 (1), 343–381.

Blume, M. E., Friend, I., 1975. The asset structure of individual portfolios and some implications for utility functions. Journal of Finance 30 (2), 585–603.

Blume, M. E., Keim, D. B., 2012. Institutional investors and stock market liquidity: trends and relationships. Jacobs Levy Equity Management Center for Quantitative Financial Research Paper.

Blume, M. E., Keim, D. B., 2014. The changing nature of institutional stock investing. Critical Finance Review 6, 1–41.

Boehmer, E., Jones, C. M., Zhang, X., 2013. Shackling short sellers: The 2008 shorting ban. Review of Financial Studies 26 (6), 1363–1400.

Bouchaud, J.-P., Bonart, J., Donier, J., Gould, M., 2018. Trades, quotes and prices: financial markets under the microscope. Cambridge University Press.

Brainard, W. C., Tobin, J., 1968. Pitfalls in financial model building. The American Economic Review 58 (2), 99–122.

Brandt, M. W., Santa-Clara, P., 2006. Dynamic portfolio selection by augmenting the asset space. Journal of Finance 61 (5), 2187–2217.

Brandt, M. W., Santa-Clara, P., Valkanov, R., 2009. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. Review of Financial Studies 22 (9), 3411–3447.

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

Buffa, A. M., Hodor, I., 2018. Institutional investors, heterogeneous benchmarks and the comovement of asset prices. Working Paper.

Campbell, J. Y., 2017. Financial decisions and markets: a course in asset pricing. Princeton University Press.

Campbell, J. Y., Cochrane, J. H., 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. Journal of Political Economy 107 (2), 205–251.

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., 2012. The econometrics of financial markets. Princeton University Press.

Campbell, J. Y., Shiller, R. J., 1988b. Stock prices, earnings, and expected dividends. Journal of Finance 43 (3), 661–676.

Cardell, N. S., 1997. Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. Econometric Theory.

Carhart, M. M., 1997. On persistence in mutual fund performance. Journal of Finance 52 (1), 57–82.

Chang, Y. C., Hong, H., Liskovich, I., 2015. Regression discontinuity and the price effects of stock market indexing. The Review of Financial Studies 28 (1), 212–246.

Cheng, X., Dou, W. W., Liao, Z., 2022. Macro-finance decoupling: Robust evaluations of macro asset pricing models. Econometrica 90 (2), 685–713.

Cornell, B., Roll, R., 2005. A delegated-agent asset-pricing model. Financial Analysts Journal 61 (1), 57–69.

Cremers, K. M., Petajisto, A., 2009. How active is your fund manager? a new measure that predicts performance. Review of Financial Studies 22 (9), 3329–3365.

Cuoco, D., Kaniel, R., 2011. Equilibrium prices in the presence of delegated portfolio management. Journal of Financial Economics 101 (2), 264–296.

Da, Z., Larrain, B., Sialm, C., Tessada, J., 2018. Destabilizing financial advice: Evidence from pension fund reallocations. The Review of Financial Studies 31 (10), 3720–3755.

Daniel, K., Grinblatt, M., Titman, S., Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. Journal of Finance 52 (3), 1035–1058.

Daniel, K. D., Klos, A., Rottke, S., 2022. The dynamics of disagreement. Review of Financial Studies.

Deaton, A., Muellbauer, J., 1980. An almost ideal demand system. The American Economic Review 70 (3), 312–326.

Domencich, T. A., McFadden, D., 1975. Urban travel demand-a behavioral analysis. North-Holland, Amsterdam.

Dou, W. W., Ji, Y., Wu, W., 2021. Competition, profitability, and discount rates. Journal of Financial Economics 140 (2), 582–620.

Dou, W. W., Kogan, L., Wu, W., Forthcoming. Common fund flows: Flow hedging and factor pricing. Journal of Finance.

Drechsler, I., Drechsler, Q. F., 2014. The shorting premium and asset pricing anomalies. (No. w20282). National Bureau of Economic Research Working Paper.

Drechsler, I., Savov, A., Schnabl, P., 2018. A model of monetary policy and risk premia. Journal of Finance 73 (1), 317–373.

Dubin, J. A., 1986. A nested logit model of space and water heat system choice. Marketing Science.

Falaris, E. M., 1987. A nested logit migration model with selectivity. International Economic Review.

Fama, E. F., French, K. R., 1992. The cross-section of expected stock returns. Journal of Finance 47 (2), 427–465.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116 (1), 1–22.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81 (3), 607–636.

Forinash, C. V., Koppelman, F. S., 1993. Application and interpretation of nested logit models of intercity mode choice. Transportation research record 1413.

Frazzini, A., Israel, R., Moskowitz, T. J., 2018. Trading costs. Working Paper.

Frazzini, A., Pedersen, L. H., 2014. Betting against beta. Journal of Financial Economics 111 (1), 1–25.

Gabaix, X., Koijen, R. S., 2020. Granular instrumental variables. (No. w28204). National Bureau of Economic Research Working Paper.

Gabaix, X., Koijen, R. S., 2021. In search of the origins of financial fluctuations: The inelastic markets hypothesis. (No. w28967). National Bureau of Economic Research Working Paper.

Gabaix, X., Koijen, R. S., Mainardi, F., Oh, S., Yogo, M., 2022. Asset demand of us households. Working Paper.

Gandhi, A., Nevo, A., 2021. Empirical models of demand and supply in differentiated products industries. Handbook of Industrial Organization 4 (1), 63–139.

Geczy, C. C., Musto, D. K., Reed, A. V., 2002. Stocks are special too: An analysis of the equity lending market. Journal of Financial Economics 66 (2-3), 241–269.

Giglio, S., Xiu, D., 2021. Asset pricing with omitted factors. Journal of Political Economy 129 (7), 1947–1990.

Goetzmann, W. N., Kumar, A., 2008. Equity portfolio diversification. Review of Finance 12 (3), 433–463.

Goldman, E., Slezak, S. L., 2003. Delegated portfolio management and rational prolonged mispricing. Journal of Finance 58 (1), 283–311.

Gompers, P. A., Metrick, A., 2001. Institutional investors and equity prices. Quarterly Journal of Economics 116 (1), 229–259.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33 (5), 2223–2273.

Hansen, L. P., Heaton, J., Yaron, A., 1996. Finite-sample properties of some alternative gmm estimators. Journal of Business and Economic Statistics 14 (3), 262–280.

Hartzmark, S. M., Solomon, D. H., 2022. Predictable price pressure. Working paper.

Haruvy, E., Noussair, C. N., 2006. The effect of short selling on bubbles and crashes in experimental spot asset markets. Journal of Finance 61 (3), 1119–1157.

He, Z., Kelly, B., Manela, A., 2017. Intermediary asset pricing: New evidence from many asset classes. Journal of Financial Economics 126 (1), 1–35.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28 (3), 650–705.

Hugonnier, J., Kaniel, R., 2010. Mutual fund portfolio choice in the presence of dynamic flows. Mathematical Finance 20 (2), 187–227.

Ivkovic, Z., Sialm, C., Weisbenner, S., 2008. Portfolio concentration and the performance of individual investors. Journal of Financial and Quantitative Analysis 43 (3), 613–655.

Kacperczyk, M., Nieuwerburgh, S. V., Veldkamp, L., 2014. Time-varying fund manager skill. Journal of Finance 69 (4), 1455–1484.

Kan, R., Zhang, C., 1999a. Gmm tests of stochastic discount factor models with useless factors. Journal of Financial Economics 54 (1), 103–127.

Kan, R., Zhang, C., 1999b. Two-pass tests of asset pricing models with useless factors. Journal of Finance 54 (1), 203–235.

Kaniel, R., Kondor, P., 2013. The delegated lucas tree. Review of Financial Studies 26 (4), 929–984.

Koijen, R. S., 2014. The cross-section of managerial ability, incentives, and risk preferences. Journal of Finance 69 (3), 1051–1098.

Koijen, R. S., Koulischer, F., Nguyen, B., Yogo, M., 2021. Inspecting the mechanism of quantitative easing in the euro area. Journal of Financial Economics 140 (1), 1–20.

Koijen, R. S., Richmond, R. J., Yogo, M., 2019. Which investors matter for global equity valuations and expected returns? Working paper.

Koijen, R. S., Yogo, M., 2019. A demand system approach to asset pricing. Journal of Political Economy 127 (4), 1475–1515.

Koijen, R. S., Yogo, M., 2020. Exchange rates and asset prices in a global demand system. (No. w27342). National Bureau of Economic Research Working Paper.

Krishnamurthy, A., Vissing-Jorgensen, A., 2007. The demand for treasury debt. (No. 12881). National Bureau of Economic Research Working Paper.

Lakonishok, J., Shleifer, A., Vishny, R., 1994. Contrarian investment, extrapolation, and risk. Journal of Finance 49 (5), 1541–1578.

Lou, D., 2012. A flow-based explanation for return predictability. The Review of Financial Studies 25 (12), 3457–3489.

Lucas, R. E., 1978. Asset prices in an exchange economy. Econometrica 46 (6), 1429–1445.

Lynch, A. W., 2001. Portfolio choice and equity characteristics: Characterizing the hedging demands induced by return predictability. Journal of Financial Economics 62 (1), 67–130.

McFadden, D., 1981. Econometric models of probabilistic choice. Structural analysis of discrete data with econometric applications 198272.

Novy-Marx, R., 2013. The other side of value: The gross profitability premium. Journal of Financial Economics 108 (1), 1–28.

Ofek, E., Richardson, M., 2003. Dotcom mania: The rise and fall of internet stock prices. Journal of Finance 58 (3), 1113–1137.

Pastor, L., Stambaugh, R. F., Taylor, L. A., 2020. Fund tradeoffs. Journal of Financial Economics 138 (3), 614–634.

Pollet, J. M., Wilson, M., 2008. How does size affect mutual fund behavior? Journal of Finance 63 (6), 2941–2969.

Rosen, S., 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy 82 (1), 34–55.

Shleifer, A., 1986. Do demand curves for stocks slope down? Journal of Finance 41 (3), 579–590.

Stambaugh, R. F., Yu, J., Yuan, Y., 2012. The short of it: Investor sentiment and anomalies. Journal of Financial Economics 104 (2), 288–302.

Tang, Y., Whitelaw, R. F., 2011. Time-varying sharpe ratios and market timing. Quarterly Journal of Finance 1 (3), 465–493.

Tobin, J., 1969. A general equilibrium approach to monetary theory. Journal of Money, Credit and Banking 1 (1), 15–29.

Vayanos, D., Woolley, P., 2013. An institutional theory of momentum and reversal. Review of Financial Studies 26 (5), 1087–1145.

Whitelaw, R. F., 1994. Time variations and covariations in the expectation and volatility of stock market returns. Journal of Finance 49 (2), 515–541.

Yu, J., 2007. Closed-form likelihood approximation and estimation of jump-diffusions with an application to the realignment risk of the chinese yuan. Journal of Econometrics 141 (2), 1245–1280.

Table 1: Estimated Coefficients on Diversification Value

	Long Investors	Short Sellers
Estimate	0.06	0.31
S.E.	(0.01)	(0.15)

This table reports the estimated coefficients θ on diversification value defined in Equations 4 and 8. Standard errors are reported in parentheses.

Table 2: Standard Deviation and Auto-Correlation of Aggregate Latent Demand

Investor Type	Standard Deviation	Lag 1 Auto-Correlation	S.E. of Auto-Correlation
Banks	0.38	0.38	(0.01)
Insurance companies	0.48	0.13	(0.01)
Investment advisors	0.42	0.22	(0.00)
Mutual funds	0.53	0.46	(0.00)
Pension funds	0.41	0.39	(0.01)
Households	0.45	0.17	(0.08)
Short sellers	0.17	0.43	(0.07)

This table reports the standard deviation and lag 1 auto-correlation of the approximate aggregate latent demand in Equation 20 by investor type. Standard errors of auto-correlation are reported in parentheses. The quarterly sample period is from 1981q1 to 2019q4.

Table 3: Variance Decomposition of Stock Returns

	Full Sample % of Variance	Post 2000 % of Variance
Supply-side:		
Shares outstanding	1.0	1.1
	(0.1)	(0.2)
Stock characteristics	9.2	8.9
	(0.3)	(0.5)
Dividend yield	0.5	0.3
	(0.0)	(0.0)
Demand-side:		
Total Wealth	2.5	3.8
	(0.2)	(0.4)
Aggregate coefficients	0.1	0.2
	(0.1)	(0.1)
Coefficients on characteristics	3.3	3.9
	(0.3)	(0.6)
Aggregate latent demand	0.5	1.1
	(0.2)	(0.3)
Asset-specific latent demand: Extensive margin	14.0	13.1
	(0.2)	(0.4)
Asset-specific latent demand: Intensive margin	69.0	67.7
	(0.4)	(0.6)
Observations	154,343	82,051

This table reports the cross-sectional variance decomposition of annual stock returns into supply- and demand-side effects. Heteroskedasticity-robust standard errors are reported in parentheses. The full sample period is from 1982 to 2019.

Table 4: Institutional Demand and Return Premiums

Return Premium (in %)	Real Data		Counter Factual	
Panel A: No institutional demand				
Size	-20.18	(6.55)	-0.73	(7.32)
Value	18.69	(5.94)	11.89	(3.33)
Profitability	5.37	(2.70)	5.40	(2.50)
Investment	-23.00	(4.53)	-1.24	(3.35)
Panel B: No shorting				
Size	-20.18	(6.55)	-15.54	(5.45)
Value	18.69	(5.94)	15.98	(5.29)
Profitability	5.37	(2.70)	5.31	(2.26)
Investment	-23.00	(4.53)	-17.71	(3.88)

This table reports the the average annual return spreads between the top decile and the bottom decile sorted by size, value, profitability, and investment in real data and counter-factual scenarios. In Panel A, institutions are neutral towards the corresponding characteristics. In Panel B, short sales are banned. Standard errors are reported in parentheses. The annual sample period is from 1982 to 2019.

Table 5: Repricing Effect of a Short-Sale Ban

Market Cap Decile	Average Repricing (in %)		Average Repricing (in %)		BE/ME Ratio Decile	Average Repricing (in %)		Average Repricing (in %)	
	08q3-09q4		Full Sample			08q3-09q4		Full Sample	
1	-1.04	(0.13)	-0.51	(0.09)	1	3.31	(0.33)	1.50	(0.13)
2	-0.54	(0.23)	-0.40	(0.09)	2	2.71	(0.41)	1.00	(0.08)
3	0.76	(0.54)	0.31	(0.16)	3	3.81	(0.41)	1.31	(0.11)
4	3.37	(1.20)	1.48	(0.27)	4	2.45	(0.48)	1.13	(0.09)
5	9.30	(2.57)	2.83	(0.38)	5	2.15	(0.45)	1.09	(0.10)
6	16.18	(2.99)	4.35	(0.51)	6	2.52	(0.77)	0.92	(0.08)
7	19.24	(2.94)	5.27	(0.55)	7	2.72	(0.68)	0.93	(0.08)
8	18.23	(2.03)	4.32	(0.44)	8	2.65	(0.47)	1.06	(0.10)
9	9.20	(0.39)	3.13	(0.24)	9	4.62	(0.85)	1.13	(0.11)
10	0.34	(0.17)	0.40	(0.03)	10	3.35	(0.35)	2.08	(0.17)

This table reports the average repricing defined in Equation 28 around the 2008 financial crisis and in the full sample on stocks sorted by market capitalization and book-to-market ratio. Standard errors are reported in parentheses. The quarterly sample period is from 1981q1 to 2019q4.

Table 6: Repricing Effect of a Short-Sale Ban - Continued

Profitability Decile	Average Repricing (in %)		Average Repricing (in %)		Investment Decile	Average Repricing (in %)		Average Repricing (in %)	
	08q3-09q4		Full Sample			08q3-09q4		Full Sample	
1	13.58	(3.05)	4.96	(0.61)	1	7.55	(1.65)	1.87	(0.21)
2	7.48	(0.88)	3.13	(0.39)	2	6.52	(1.13)	1.64	(0.16)
3	6.31	(0.67)	1.78	(0.18)	3	4.01	(1.05)	1.19	(0.11)
4	4.39	(0.54)	1.23	(0.11)	4	2.55	(0.73)	0.84	(0.08)
5	3.30	(0.53)	1.03	(0.09)	5	2.45	(0.41)	0.82	(0.07)
6	3.26	(0.89)	1.00	(0.10)	6	2.01	(0.59)	0.85	(0.07)
7	2.95	(0.66)	0.85	(0.09)	7	1.19	(0.40)	0.84	(0.08)
8	2.06	(0.42)	0.93	(0.08)	8	2.44	(0.63)	1.16	(0.10)
9	1.57	(0.32)	0.82	(0.06)	9	3.69	(0.30)	1.60	(0.13)
10	2.69	(0.36)	1.26	(0.09)	10	5.37	(0.73)	2.11	(0.22)

This table reports the average repricing defined in Equation 28 around the 2008 financial crisis and in the full sample on stocks sorted by profitability and investment. Standard errors are reported in parentheses. The quarterly sample period is from 1981q1 to 2019q4.

Table 7: Return Predictability of Dividend-Price Ratio

Horizon	h=1	h=2	h=3	h=1	h=2	h=3
	Real Data			Idiosyncratic Latent Demand		
D/P Coef	0.03	0.07	0.11	0.00	0.04	0.05
P-value	0.09	0.01	0.00	0.87	0.19	0.20
	Neutral Towards Sharpe			Combined Effects		
D/P Coef	0.03	0.07	0.11	0.00	0.04	0.05
P-value	0.16	0.04	0.01	0.93	0.30	0.24

This table reports coefficient estimates and the corresponding p-values from regressions of future log excess return on log dividend-price ratio for the value-weighted market portfolio. The top-left panel presents results from real data, and the other three panels present results from simulated data. The sample period for quarterly regressions is from 1981q2 to 2019q4.

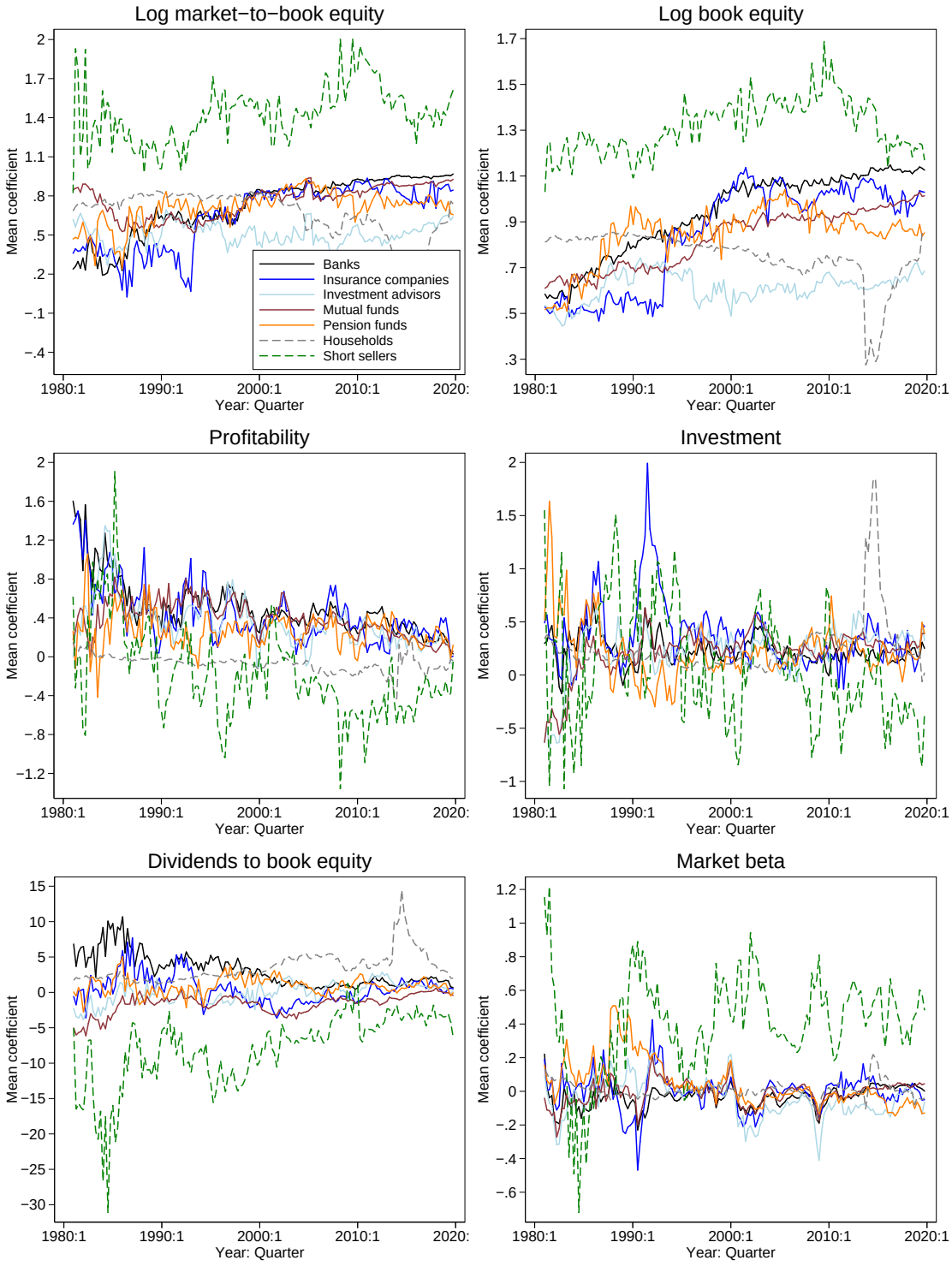

Figure 1: Estimated coefficients on characteristics. This figure reports the cross-sectionally averaged coefficients in Equation 14 by investor type, weighted by total investment in risky assets. The quarterly sample period is from 1981q1 to 2019q4.

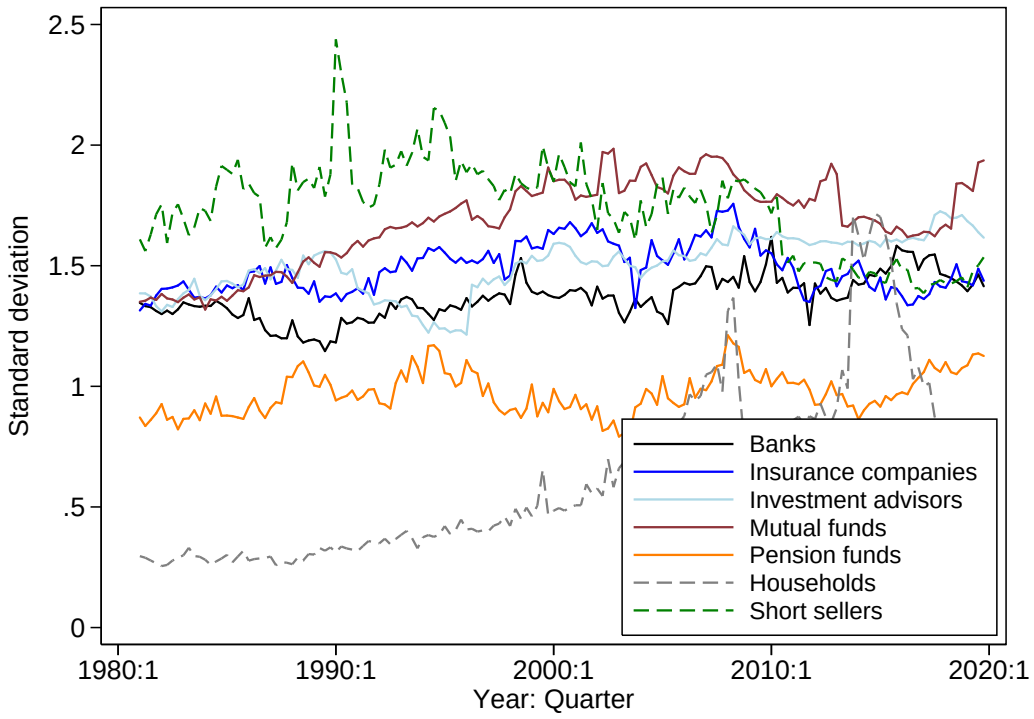

Figure 2: Standard deviation of asset-specific latent demand. This figure reports the cross-sectional standard deviation of log asset-specific latent demand in Equation 14 by investor type, weighted by total investment in risky assets. The quarterly sample period is from 1981q1 to 2019q4.

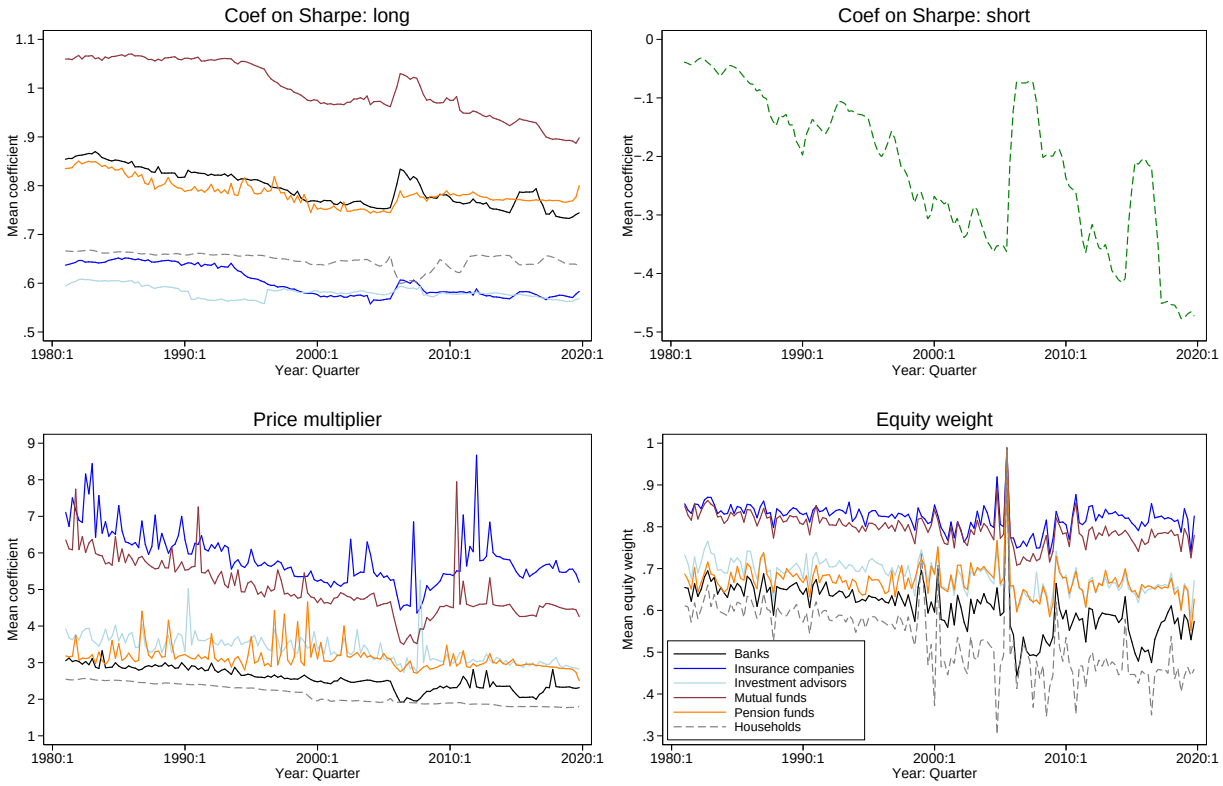

Figure 3: Estimated aggregate coefficients and equity weights. This figure reports the cross-sectionally averaged aggregate coefficients and portfolio weights on equities (risky assets) in Equations 18 and 20 by investor type, weighted by total investment in risky assets. The quarterly sample period is from 1981q1 to 2019q4.

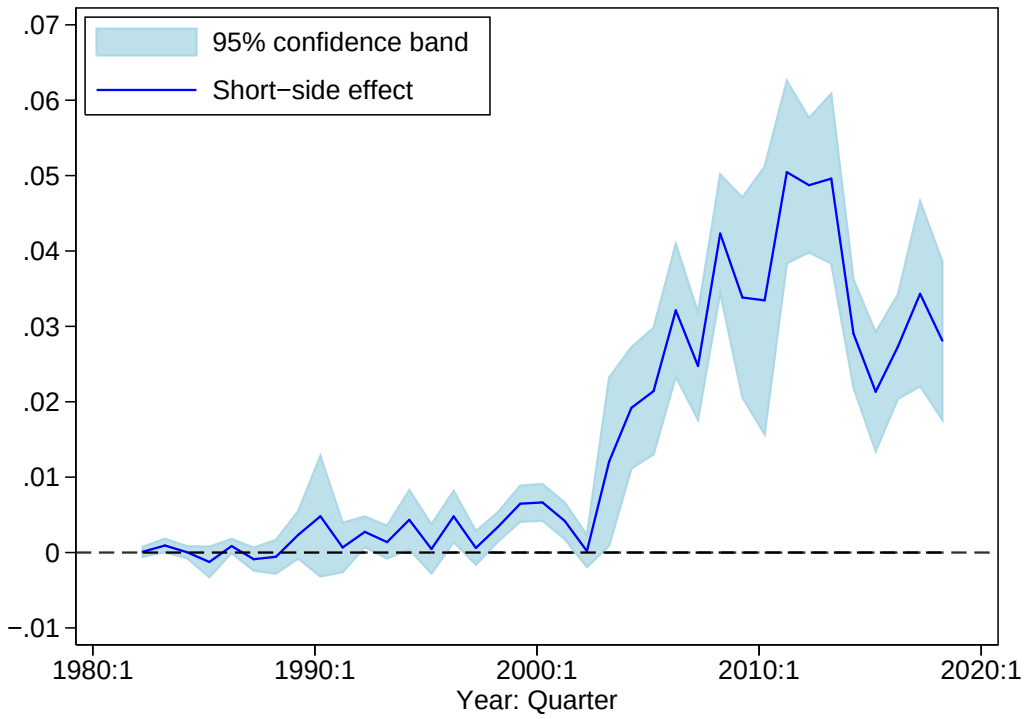

Figure 4: Variance decomposition of stock returns on short sales. The cross-sectional variance of annual stock returns is decomposed into long- and short-side effects. This figure reports the proportion of return variation explained by short sales for each year and the 95% confidence band. The annual sample period is from 1982 to 2019.

Figure 5: S&P 500 index and overall index with counter-factual short-selling activities. Overall index consists of all stocks in the sample. The counter-factual index values are computed by banning or increasing short sales. The quarterly sample period displayed is from 1996q1 to 2003q4.

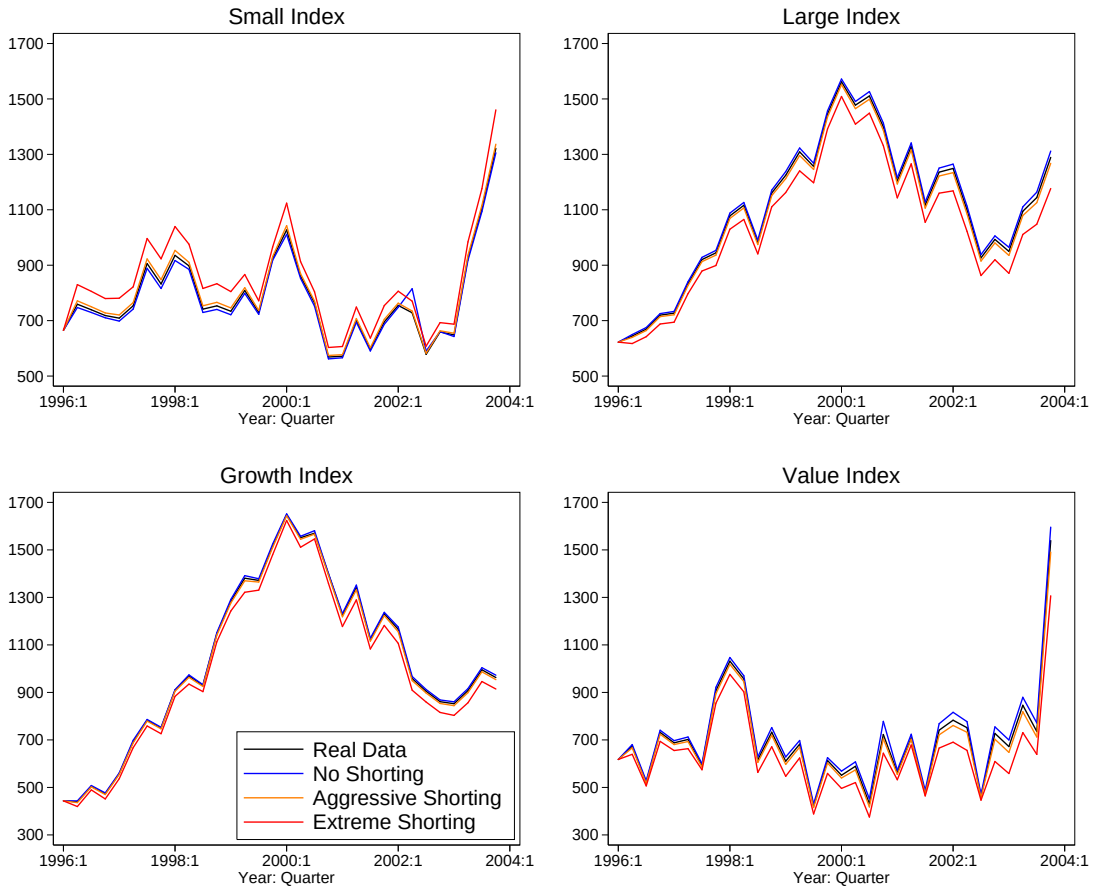

Figure 6: Different stock indices with counter-factual short-selling activities. Small and large indices consist of the lower 30% or upper 30% of stocks sorted by market capitalization. Growth and value indices consist of the lower 30% or upper 30% of stocks sorted by book-to-market ratio. The counter-factual index values are computed by banning or increasing short sales. The quarterly sample period displayed is from 1996q1 to 2003q4.

A Data Construction Details

I summarize the detailed data construction in this section.

A.1 Asset-Level Data

The data on stock prices, dividends, returns, and shares outstanding are from the Center for Research in Security Prices (CRSP) Monthly Stock Database. In cases of missing prices or shares outstanding, I supplement these data using the Thomson Reuters Institutional Holdings Database (s34 file), if available. Eventually, the sample is restricted to assets with non-missing prices and shares outstanding. The data on fundamentals are from the Compustat North America Fundamentals Annual and Quarterly Databases. Based on the CRSP/Compustat Merged (CCM) link table, the CRSP data are merged with the most recent Compustat data as of at least 6 months and no more than 18 months prior to the trading date. Finally, the data on risk-free rate and market excess return are from the Kenneth R. French Data Library.

Asset characteristics include log market equity, log book equity, profitability, investment, dividends to book equity, and market beta. Market equity is price per share times the number of shares outstanding. Book equity is stockholders' equity plus deferred taxes and investment tax credit minus preferred stock. Profitability is the ratio of operating profits to book equity, where operating profits is computed as revenue minus the sum of cost of goods sold, selling, general and administrative expenses, and interest expenses. Investment is the annual log growth rate of total assets. Dividends to book equity is the ratio of annual dividends per split-adjusted share times shares outstanding to book equity. Market beta is estimated via a 60-month rolling window regression of monthly excess return onto market excess return, with a minimum window of 24 months. To remove extreme outliers, I winsorize profitability, investment, and market beta at 2.5% and 97.5% and the dividends to book equity ratio at 97.5% for each time period.

A.2 Aggregate-Level Data

The data on risk-free rate and market excess return are from the Kenneth R. French Data Library. Other aggregate data used in the construction of aggregate Sharpe ratio are from the U.S. department of treasury and Federal Reserve Economic Data (FRED).

The aggregate Sharpe ratio is constructed following Whitelaw (1994) and Tang and Whitelaw (2011). A GARCH(1,1) model is estimated over monthly market excess return

from 1953m4 to 2019m12. In the mean equation, I include the Baa-Aaa spread, the dividend yield, the one-year Treasury yield and the lagged market excess return. In the variance equation, I include the one-year Treasury yield and the commercial paper-Treasury spread. Once the model is estimated, the monthly Sharpe ratio is computed as the ratio of expected market excess return over the corresponding standard deviation in the current month. Finally, the quarterly Sharpe ratio is computed as the average of the monthly Sharpe ratio multiplied by $\sqrt{3}$ within the quarter.

A.3 Investor-Level Data

The data on institutional holdings are from the Thomson Reuters Institutional Holdings Database (s34 file), with a coverage from 1980q1 to 2019q4. The stock-level short interest data are from Compustat North America Supplemental Short Interest File. The holdings data are merged with CRSP-Compustat data by CUSIP number.

The data on institution types are either from Koijen and Yogo (2019), who have hand corrected noticeable errors in the type data from the Thomson Reuters Institutional Holdings Database (s34 file), or directly from the latter if the former is unavailable. According to the data, institutions are grouped into six types: banks, insurance companies, investment advisors, mutual funds, pension funds, and other 13F institutions. I define the household sector as the investor who holds the residual shares between total shares outstanding and the sum of shares held by the institutions and short sellers. I also include in the household sector any institution with less than $10 million of total investment in risky assets, no base risky asset, or no risky assets other than the base asset in the investment universe. Therefore, the household sector represents direct household holdings and small institutional investors.

For each investor, total investment in risky assets $A^*_{i,t}$ is computed as the total market value of asset holdings reported in Form 13F. The effective number of shares of the aggregate risky asset market held by each investor is computed as the total investment in risky assets divided by the aggregate price of the risky asset class. The aggregate price is determined by Equation 1, with the divisor set at 1 million. Following Koijen and Yogo (2019), I define the investment universe for each investor $N_{i,t}$ as all risky assets that are currently held or were ever held in the previous 11 quarters. The conditional portfolio weights within the risky asset class $w^{n|*}_{i,t}$ is the market value of individual asset holdings over total investment in risky assets. For the investor-level characteristics used in Section 4.2, I measure investor size with the log of average past investment in risky assets, $\log(1 + \frac{1}{3}\sum_{k=1}^{3} |A^*_{i,t-k}|)$, and measure the activeness of management with average past active share, $\frac{1}{3}\sum_{k=1}^{3} \text{actshr}_{i,t-k}$. An investor's

active share is computed as

$$\text{actshr}_{i,t} = \frac{1}{2} \sum_{n \in N_{i,t}} \left| w_{i,t}^{n|*} - w_{i,t}^{n,me|*} \right|$$

where $w_{i,t}^{n,me|*}$ is the market-weighted risky portfolio within investor i's investment universe at time t.

Table 8 summarizes the 13F institutions as a whole and by type from 1980q1 to 2019q4. At the beginning of the sample, 535 institutions managed 34 percent of the stock market. This number grows steadily to 4,051 institutions that managed 67 percent of the stock market by the end of the sample. From 2015 to 2019, the median institution managed $385 million and held 70 stocks, while larger institutions at the 90th percentile managed $5,747 million and more diversified portfolios at the 90th percentile consisted of 503 stocks.

Table 8: Summary of 13F Institutions by Type

Period	Number of institutions	% of market held	Total investment in risky assets ($ in mil)		Number of stocks held		Number of stocks in investment universe	
			Median	90th percentile	Median	90th percentile	Median	90th percentile
				All Institutions				
1980–1984	535	34	347	2,740	123	401	191	544
1985–1989	768	42	418	3,712	123	471	222	731
1990–1994	964	47	426	4,792	113	541	207	863
1995–1999	1,317	52	493	6,936	109	610	191	1,032
2000–2004	1,798	58	399	6,356	97	579	183	1,100
2005–2009	2,478	66	378	5,810	83	521	163	1,049
2010–2014	2,986	65	370	5,777	74	498	139	928
2015–2019	4,051	67	385	5,747	70	503	125	888
				Banks				
1980–1984	204	14	340	2,925	164	516	243	689
1985–1989	202	15	506	4,342	210	631	337	944
1990–1994	199	13	502	6,389	215	767	341	1,156
1995–1999	173	11	630	16,679	238	1,177	360	1,806
2000–2004	159	11	469	22,432	234	1,410	379	2,202
2005–2009	157	11	447	18,736	210	1,452	339	2,321
2010–2014	154	10	515	19,092	192	1,291	306	1,989
2015–2019	156	11	807	33,341	229	1,693	346	2,196
				Insurance companies				
1980–1984	58	3	396	2,322	100	392	160	523
1985–1989	65	3	473	2,663	104	443	209	692
1990–1994	68	3	656	3,628	126	585	246	886
1995–1999	68	4	1,336	8,416	162	1,013	312	1,421
2000–2004	57	4	1,469	13,023	240	1,830	459	2,256
2005–2009	46	4	1,765	29,212	344	2,027	542	2,583
2010–2014	44	2	1,492	38,266	242	2,039	411	2,434
2015–2019	48	2	2,325	52,203	197	2,305	315	2,637
				Investment advisors				
1980–1984	134	5	283	1,233	88	239	151	380
1985–1989	262	8	262	1,301	78	241	153	498
1990–1994	362	9	225	1,385	75	232	143	464
1995–1999	657	7	280	1,485	76	223	131	451
2000–2004	1,138	9	284	1,806	77	247	146	539
2005–2009	1,819	16	302	2,647	70	297	141	661
2010–2014	2,388	19	309	3,098	63	306	122	632
2015–2019	3,349	20	330	3,241	63	353	113	666
				Mutual funds				
1980–1984	91	8	533	3,573	150	408	244	559
1985–1989	180	12	708	5,110	140	464	282	779
1990–1994	280	18	950	6,793	138	545	272	942
1995–1999	364	27	1,694	16,317	157	745	313	1,347
2000–2004	323	30	2,463	26,588	194	1,222	421	2,006
2005–2009	271	32	3,114	48,075	208	1,172	452	2,111
2010–2014	244	29	3,815	46,161	201	1,134	412	2,017
2015–2019	218	30	5,064	61,574	209	1,300	425	2,048
				Pension funds				
1980–1984	24	3	1,213	3,873	114	401	151	479
1985–1989	32	4	1,222	7,682	236	682	316	791
1990–1994	33	4	1,036	15,378	308	994	472	1,215
1995–1999	30	3	1,922	27,552	425	1,353	701	1,631
2000–2004	35	3	4,535	39,489	620	2,092	939	2,534
2005–2009	40	3	6,451	37,865	716	2,312	1,092	2,709
2010–2014	52	2	4,951	27,082	549	1,698	796	2,294
2015–2019	53	2	7,167	38,118	584	1,617	870	2,203
				Other				
1980–1984	24	1	225	1,517	69	189	95	265
1985–1989	28	1	255	1,269	69	228	104	434
1990–1994	21	1	262	2,494	75	169	112	384
1995–1999	23	0	293	1,998	84	141	120	377
2000–2004	87	0	191	1,656	53	259	100	427
2005–2009	145	1	197	2,718	41	333	99	743
2010–2014	104	2	280	7,252	48	640	105	1,075
2015–2019	227	2	333	5,817	38	494	71	781
				Short sellers				
1980–1984	1	0	-1,724	-1,724	717	717	974	974
1985–1989	1	0	-6,609	-6,609	862	862	1,121	1,121
1990–1994	1	0	-16,820	-16,820	1,071	1,071	1,277	1,277
1995–1999	1	-1	-72,631	-72,631	1,515	1,515	1,658	1,658
2000–2004	1	-1	-157,100	-157,100	2,077	2,077	2,217	2,217
2005–2009	1	-2	-407,932	-407,932	3,603	3,603	3,619	3,619
2010–2014	1	-2	-564,554	-564,554	3,326	3,326	3,328	3,328
2015–2019	1	-2	-890,730	-890,730	3,480	3,480	3,482	3,482

This table reports the time-series mean of each summary statistic within the given period. The quarterly sample period is from 1980q1 to 2019q4.

B Simulation of Aggregate Estimation

I demonstrate the performance of the aggregate estimation procedure proposed in Section 4.2 in a simple simulation environment. To speed up each estimation, I set the number of periods at $T = 500$, the number of individual risky assets at $N = 11$, the number of investors at $I = 50$, and all coefficients to be constant instead of time-varying. For simplicity, I drop households and short sellers and independently draw investors' total wealth $A_{i,t}$ from $N(500, 5)$.

The conditional portfolio weights within the risky asset class are generated as in Equations 2 and 3. Specifically,

$$\frac{w_{i,t}^{n|*}}{w_{i,t}^{1|*}} = \exp\left(\beta_0 + \beta_1 me_t^n + \beta_2 x_t^n\right)\xi_{i,t}^n, \qquad n = 2, ..., N$$

where the coefficients are $\beta_0 = -1,\ \beta_1 = 0,\ \beta_2 = 1$, the characteristics are drawn from $x_t \overset{i.i.d.}{\sim} U[0, 0.1]$, and the latent demand is drawn from $\log \xi_{i,t}^n \overset{i.i.d.}{\sim} N(0, 0.3^2)$. To focus solely on the aggregate estimation, I assume the econometrician has knowledge of the coefficients on asset characteristics and asset-specific latent demand when constructing the instruments proposed in Equations 21–23.

The portfolio weights between asset classes are generated as in Equations 4–6. Specifically,

$$\begin{aligned} \frac{w_{i,t}^*}{w_{i,t}^0} &= \exp\left(\alpha_0 + \alpha_1 SR_t + \theta\Gamma_{i,t}^* + \tilde{\xi}_{i,t}^*\right) \\ \Gamma_{i,t}^* &= \log\left(1 + \sum_{n>1} \exp(\beta_0 + \beta_1 me_t^n + \beta_2 x_t^n)\xi_{i,t}^n\right) \end{aligned} \tag{29}$$

where the coefficients are $\alpha_0 = 1.5,\ \alpha_1 = 0.6,\ \theta = 0.1$, which are set to match the moments and range of the estimated portfolio weights on equities. The aggregate latent demand is generated through $\tilde{\xi}_{i,t}^* = \eta_t + u_{i,t}$, where $\eta_t \sim N(0, 0.5^2)$ is the common factor and $u_{i,t} \overset{i.i.d.}{\sim} N(0, 0.3^2)$ is the idiosyncratic part.

Finally, Sharpe ratio and prices are determined endogenously by market clearing 9. I then follow the exact procedure described in Section 4.2 to estimate the aggregate coefficients. Figure 7 reports the distributions of estimates from 1,000 replications. The key takeaway is that the proposed estimator uncovers aggregate coefficients accurately, with the average

estimation error strictly smaller than 0.01 for all coefficients.

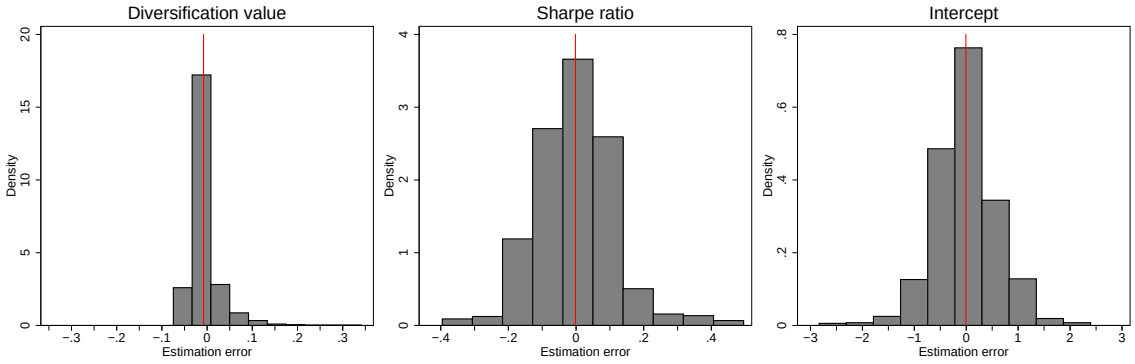

Figure 7: Simulation results of the aggregate estimation. This figure reports the distributions of estimation error from 1,000 replications for each aggregate coefficient in Equation 29. The red vertical lines indicate the average estimation error.