The Making of Momentum
A Demand-System Perspective*

Job Market Paper

Paul Huebner
UCLA

November 28, 2022
Download the latest version here

Abstract

I develop a framework to quantify which features of investors' trading strategies lead to momentum in equilibrium. Specifically, I distinguish two channels: persistent demand shocks, capturing underreaction, and the term structure of demand elasticities, representing an intensity of arbitrage activity that decreases with investor horizon. I introduce both aspects of dynamic trading into an asset demand system and discipline the model using the joint behavior of portfolio holdings and prices. I estimate the demand of institutional investors in the U.S. stock market between 1999 and 2020. On average, investors respond more to short-term than longer-term price changes: the term structure of elasticities is downward-sloping. My estimates suggest that this channel is the primary driver of momentum returns. Moreover, in the cross-section, stocks with more investors with downward-sloping term structures of elasticities exhibit stronger momentum returns by 7% per year.

*Email: paul.huebner.phd@anderson.ucla.edu. I am incredibly grateful to my advisors Valentin Haddad, Mike Chernov, Andrea Eisfeldt, Lars Lochstoer, and Stavros Panageas for their invaluable support and guidance. I am thankful for helpful comments by Edward Kim, Erik Loualiche, Denis Mokanov, Tyler Muir, Gregor Schubert, Danyu Zhang, Geoffery Zheng, and seminar participants at UCLA Anderson and WU Vienna.

1 Introduction

Momentum, the tendency for past winners to outperform past losers (Jegadeesh and Titman, 1993), is one of the most challenging anomalies to understand in stock returns.[1] While many explanations for momentum have been proposed, tests of these theories have mainly focused on the behavior of returns.[2] In this paper, I take a different approach by looking at the joint behavior of investor portfolio holdings and prices. I propose a framework to measure the dynamic trading strategies of each investor and quantify how they contribute to the making of momentum in equilibrium. Looking jointly at quantities and prices gets to the heart of how momentum is created — investors' dynamic trading — and yields new insights into who are the investors driving momentum.

I highlight two broad mechanisms that generate momentum: the persistence of demand shocks, representing relative underreaction, and a downward-sloping term structure of demand elasticities, which captures different intensities of arbitrage activity across time horizons. I introduce these mechanisms into an asset demand system in the style of Koijen and Yogo (2019) and estimate it from data on portfolio holdings. My estimates show that equilibrium momentum is primarily the result of the downward-sloping term structure of demand elasticities. Market participants respond more strongly to price changes over the most recent quarter than to longer-term variation over one year.[3] My framework also predicts higher momentum returns in stocks owned by investors with a downward-sloping term structure of elasticities. Accordingly, I sort stocks based on their aggregate term structure of elasticities and find 7% higher momentum returns in stocks where it is more steeply downward-sloping.

Which aspects of how people trade lead to momentum? The first one is, in the language of demand systems, about the persistence of demand shocks. It is the mechanism behind classic

[1] Fama (2014), in his Nobel Prize Lecture, acknowledges momentum as "the biggest challenge to market efficiency."

[2] Some notable exceptions include early work by Grinblatt, Titman, and Wermers (1995) and Grinblatt and Keloharju (2000) and more recent work by Chui, Subrahmanyam, and Titman (2022).

[3] I specifically compare quarterly and yearly horizons to align with the empirical definition of momentum formation periods (e.g., Jegadeesh and Titman, 1993). So when I say that the term structure of elasticities is downward-sloping, I mean that it is downward-sloping at these specific frequencies.

momentum explanations through underreaction to information (e.g., Chan, Jegadeesh, and Lakonishok, 1996): Upon receiving fundamental news, investors respond only partially when incorporating it into their demand. Over time, they react more and more strongly to the information, creating a drift in prices.[4] But there is another potential source of momentum, distinct from underreaction in demand shocks: differences in investors' ability to absorb shocks across horizons, the term structure of demand elasticities. For example, consider Elon Musk selling 10 million shares of Tesla to raise capital for the acquisition of Twitter (and abstract from information effects). Initially, this demand shock is absorbed predominantly by relatively higher-frequency arbitrageurs on the lookout for fast opportunities, so the price of Tesla does not decrease much. But higher-frequency traders have short investment horizons and soon turn their attention elsewhere. So they sell their Tesla shares to investors with longer horizons, for example, active mutual funds. If the higher-frequency traders are more willing to absorb the Tesla shares than the active mutual funds, then the price of Tesla stock will decrease further. More generally, when there is a mismatch in the aggregate risk-bearing capacity at the short versus the long horizon — the term structure of demand elasticities — then the equilibrium price impact of a demand shock will increase over time. In other words, when short-run arbitrage exceeds long-run arbitrage, the term structure of demand elasticities is downward-sloping, and momentum arises.

To quantify the importance of these two channels, I incorporate the term structure of elasticities into an asset demand system in the tradition of Koijen and Yogo (2019). Introducing dynamics into a demand system leads to new challenges for identification, especially in separating the two explanations for momentum. The inclusion of price changes at different horizons creates a dynamic simultaneity problem, resulting from the combination of persistent demand shocks with the classic simultaneity problem of prices and demand. In other words, it is difficult to disentangle the dynamics of demand shocks from the evolution of

[4] The literature has put forward many foundations for such underreaction. I use the term "underreaction" in a broad sense to capture relative patterns in beliefs across time, encompassing models of delayed overreaction alongside underreaction stricto sensu. Section 2.5 summarizes these theories and shows how they generate persistence in investor demand.

investors' equilibrium responses to said shocks across time. Starting from the idea of mutual fund flow-induced trading (Lou, 2012) — facing outflows, mutual funds scale down their existing holdings to meet redemptions, thereby putting downward price pressure on the stocks they hold — I show how to construct appropriate instruments for recent and longer-term price changes to overcome the dynamic simultaneity issue. However, the relation between mutual fund flows and past fund returns, retail investors chasing fund performance, threatens exogeneity.[5] To account for it, I orthogonalize mutual-fund flows to past fund returns and past fund flows.

I estimate the model for institutional investors in the U.S. stock market between 1999 and 2020. My estimates suggest that, on average, the term structure of elasticities is downward-sloping: The market is 25% less elastic in its response to price movements over the past year compared to the past quarter. To put this number into context, consider homogenous investors with an elasticity of 4 to returns over the most recent quarter but a lower elasticity of 3 to longer-term variation at the horizon of a year. Here, investors are $(4 - 3)/4 = 25\%$ less elastic at longer horizons, so the term structure of elasticities is downward-sloping. How does a \$100 inflow affect prices? Initially, the recent elasticity of 4 implies that the extra \$100 raise the value of the stock by $\$100/4 = \$25$. Subsequently, driven by the downward-sloping term structure of elasticities, the price impact rises to $\$25/(1 - 25\%) \approx \$33$.

There is substantial variation in elasticity estimates across investors. In particular, my estimates identify a group of investors who are very active at a quarterly horizon but less so in the long run. These investors drive much of the overall pattern of downward-sloping term structures. And because of cross-sectional variation in how much they own, I find substantial variation in the slope of the aggregate term structure of elasticities across stocks as well. It is more strongly decreasing in stocks that are unprofitable, small, or have a high dividend yield.

A distinct advantage of the demand-system approach is that it is an equilibrium frame-

[5] The flow-performance relation between mutual fund flows and past fund returns was originally documented in Ippolito (1992), Chevalier and Ellison (1997), and Sirri and Tufano (1998).

work. That is, it ensures that observed prices are the equilibrium of the individual behavior of all investors. In particular, this allows me to decompose the evolution of momentum returns into components representing dynamic trading against prices, fundamentals, and demand shocks.

The downward-sloping term structure of elasticities is the primary driver of momentum returns. On its own, this phenomenon would create annualized momentum returns of about 24% between 1999 and 2020. More specifically, if investors had not changed their demand from period to period for any reason other than the term structure of elasticities, then the equilibrium-implied period-to-period price changes would have resulted in annualized momentum returns of 24%. In contrast, investor demand shocks mean-revert, creating reversal rather than momentum.[6] This observation is at odds with theories that generate momentum through underreaction. But it does not mean that underreaction to news does not exist. First, it might have played a less dominant role only recently, which is in line with ideas of momentum anomaly attenuation (Chordia, Subrahmanyam, and Tong, 2014) and the overall poor performance of classic momentum strategies between 1999 and 2020. Second, underreaction might occur under specific conditions. For example, I find that past latent demand predicts future stock fundamentals that enter investors' demand functions, consistent with Novy-Marx (2015).

I use the model estimates to design a demand-system-boosted momentum strategy. In particular, the model predicts larger momentum returns in stocks with steeply downward-sloping term structures of elasticities. Accordingly, I sort stocks into two portfolios based on their term structures of elasticities. Then, within each subset, I examine the returns to a standard momentum strategy that goes long the tercile of past winners and short past losers. While the returns to a conventional momentum strategy were low at an annualized 2% between 1999 and 2020, the returns to momentum among stocks with more steeply decreasing term structures of elasticities were higher by 7% than among stocks with flatter

[6] Similarly, Koijen and Yogo (2019) generate a profitable reversal strategy based on the mean-reversion of demand shocks.

term structures. This difference cannot be attributed to common risk factors, including the momentum factor itself, and is robust to controlling for stock size. Interestingly, momentum among stocks with steeply decreasing term structures avoids momentum crashes that standard momentum strategies experience following stock market crashes (Daniel and Moskowitz, 2016).

My results highlight the importance of incorporating both the persistence of investors' demand shocks and the downward-sloping term structure of demand elasticities into models that generate momentum in equilibrium. Most models focus on the first aspect, which captures underreaction to news by behavioral investors. However, I show this channel to be less important empirically. At the same time, existing models often ignore what my model estimates to be the primary driver of momentum: the term structure of elasticities, could reflect representing investors' differential responses to short- and long-term variation in prices. Such dynamic responses to prices could represent frictions rooted in the industrial organization of the financial industry or reflect investors' behavioral biases in processing the information contained in equilibrium prices (Bastianello and Fontanier, 2021). They are likely also important for other anomalies based on time-series patterns in prices. Most notably, my framework can be adapted to study the drivers of price reversals at short horizons below a quarter and long horizons beyond a year.[7]

Contribution to the literature. Momentum, the tendency of past winners to outperform past losers, is one of the most widely studied anomalies (Jegadeesh and Titman, 1993, 2001). It is robust: to different formation-period definitions (Grinblatt and Moskowitz, 2004; Novy-Marx, 2012), on industry, style and factor level (Moskowitz and Grinblatt, 1999; Barberis and Shleifer, 2003; Chen and De Bondt, 2004; Ehsani and Linnainmaa, 2022), across asset classes (Asness, Moskowitz, and Pedersen, 2013; Burnside, Eichenbaum, and Rebelo, 2011;

[7] The term structure might look different at such different horizons. For example, Duffie (2010) emphasizes the role of financial intermediaries' limited risk-bearing capacity at the time of a shock for the generation of short-term reversal in prices, consistent with an upward-sloping term structure of elasticities at horizons below a quarter.

Menkhoff et al., 2012), and in the time series (Moskowitz, Ooi, and Pedersen, 2012). Many mechanisms have been proposed, including both rational (Berk, Green, and Naik, 1999; Johnson, 2002; Pastor and Stambaugh, 2003; Sadka, 2006) and behavioral explanations (Long et al., 1990; Chan, Jegadeesh, and Lakonishok, 1996; Daniel, Hirshleifer, and Subrahmanyam, 1998; Barberis, Shleifer, and Vishny, 1998; Hong and Stein, 1999; Grinblatt and Han, 2005; Daniel, Klos, and Rottke, 2021). The term structure of elasticities is conceptually related to Lou and Polk (2021), who show how momentum can arise from aggregate overreaction by arbitrageurs but use the behavior of prices for measurement. A small number of papers study momentum strategies in the context of mutual funds' portfolio holdings. For example, Grinblatt, Titman, and Wermers (1995) show that mutual funds, on average, hold past winners. Dong, Kang, and Peress (2022) find that persistent but not transient flows to mutual funds predict factor-level returns because fund managers only reinvest persistent flows into factor strategies, generating factor momentum. I contribute to this literature by measuring how the dynamic trading strategies of each institutional investor make momentum in equilibrium and by emphasizing the role of arbitrage intensities across horizons.

I also contribute to the recent literature on demand systems pioneered by Koijen and Yogo (2019). Demand systems have been used to study the role of investors in the U.S. stock market (Koijen and Yogo, 2019; Koijen, Richmond, and Yogo, 2020), in an international context (Koijen and Yogo, 2020; Jiang, Richmond, and Zhang, 2020, 2022), in government- and corporate bonds (Koijen et al., 2021; Bretscher et al., 2020), and in ESG investing (Noh and Oh, 2020; van der Beck, 2021). Balasubramaniam et al. (2021) and Gabaix et al. (2022) focus on the role of households in India and the United States. Gabaix and Koijen (2020) estimate macro elasticities for the aggregate stock market. Haddad, Huebner, and Loualiche (2022) employ a demand system to study the effects of the rise of passive investing. Similar to van der Beck (2022), I identify institutions' elasticities based on their reactions to shocks from mutual funds' flow-induced trading. As a result, my elasticity estimates are higher than in static demand-based demand models (e.g., Koijen and Yogo, 2019) by a factor of about three, in line with estimates from Pavlova and Sikorskaya (2022). My key innovation to this literature is introducing the term structure of demand elasticities, which I show to be substantially downward-sloping.

Finally, I relate to the literature on segmentation in financial markets (Merton, 1987; Grossman and Miller, 1988; Shleifer and Vishny, 1997; Gromb and Vayanos, 2002; Greenwood, Hanson, and Liao, 2018). Segmentation between market participants occurs in government bonds (Guibaud, Nosbusch, and Vayanos, 2013; Greenwood and Vayanos, 2014), options (Gârleanu, Pedersen, and Poteshman, 2009), currencies (Gabaix and Maggiori, 2015), mortgage-backed securities (Gabaix, Krishnamurthy, and Vigneron, 2007), and credit default swaps (Eisfeldt et al., 2022), all asset classes in which financial intermediaries play a prominent role (Haddad and Muir, 2021). Segmentation is often the result of some form of preferred habitat (e.g., Vayanos and Vila, 2021). Siriwardane, Sunderam, and Wallen (2021) analyze segmentation in the cross-section of arbitrages. Greenwood and Vissing-Jorgensen (2018) and Jansen (2021) emphasize the role of long-term investors. I contribute to this literature by emphasizing a related but distinct form of segmentation: differences in arbitrage activity across investment horizons. This is not unlike how short-term reversal is generated through slow-moving capital (Mitchell, Pedersen, and Pulvino, 2007; Duffie, 2010), but at longer-term horizons, creating momentum rather than reversal.

2 Equilibrium Momentum from Dynamic Trading

I present an equilibrium framework for how the evolution of investor demand can lead to momentum. Two distinct mechanisms shape momentum in equilibrium: persistent demand shocks, capturing underreaction to information, and the term structure of demand elasticities representing how investors respond to price changes across horizons. I proceed by first introducing a model that incorporates both mechanisms. Then, I show how the model generates momentum and how it relates to canonical foundations of momentum from the literature.

2.1 Framework

I introduce the model of this section. There are three investors who choose how much to buy of a single asset in fixed supply S.[8] The short-term investor ST and long-term investor LT decide their demand based on the short-term return signal P_t/P_{t-1} and past long-term return signal P_{t-1}/P_{t-s} of the asset. One period corresponds to one quarter, and s captures long horizons of one year. Investor N has noisy demand, which is persistent.

These three investors play distinct roles in the model. The role of the investor with noisy demand, N, is to generate persistent demand shocks. While I do not explicitly model the source of this persistence, it is designed to capture underreaction to information shocks. The other two investors represent institutions such as mutual funds with different investment horizons, using different price signals when forming demand. More generally, institutions exist on a spectrum ranging from high-frequency traders as fast investors trading on short-term signals on the one extreme and Warren Buffet's Berkshire Hathaway as an institutional value investor on the low-frequency end of the spectrum. My investors are not placed on either extreme but live at frequencies of a quarter (ST) and a year (LT), which is to align with formation periods from momentum strategies (e.g., Jegadeesh and Titman, 1993).

I parametrize this intuition through demand functions for the three investor types. Specifically, I log-linearize demand $D(P_t/P_{t-1})$ in recent and long-term returns around zero:

$$d_t^{ST} = \underline{d}^{ST} - \mathcal{E}_{\text{recent}} \times (p_t - p_{t-1}) \tag{1}$$

$$d_t^{LT} = \underline{d}^{LT} - \mathcal{E}_{\text{long-term}} \times (p_{t-1} - p_{t-s}) \tag{2}$$

$$D_t^N = \phi \times D_{t-1}^N + \epsilon_t^N, \tag{3}$$

where lowercase letters denote log values, $p_t - p_{t-1}$ denotes the recent log return between times $t-1$ to t, and $p_{t-1} - p_{t-s}$ is the longer-term return.

[8] In the quantitative model of Section 3, I will re-introduce heterogeneity in the full cross-sections of stocks and investors.

The recent elasticity $\mathcal{E}_{\text{recent}}$ captures how aggressively the short-horizon investor ST trades against price changes over the most recent quarter. The higher $\mathcal{E}_{\text{recent}}$, the more elastic the demand of the short-horizon investor to variation in the price p_t relative to a recent reference level, p_{t-1}. That is, if prices decrease by 1% relative to the previous period, the short-horizon investor will increase her demand by $\mathcal{E}_{\text{recent}}$%. Beyond that, $\underline{d}^{ST}$ captures an average baseline demand for the short-horizon investor. This price-insensitive component can, for example, reflect preferences for the asset based on ESG criteria.

In contrast, the long-horizon investor LT only scans prices at a lower frequency, meaning they only form demand based on prices one quarter ago. The elasticity $\mathcal{E}_{\text{long-term}}$ captures how contrarian the long-term investor is, or equivalently, how elastically she trades against longer-term variation at the frequency of a year. And again, $\underline{d}^{LT}$ captures baseline demand for the investor.

Finally, the demand of investor N includes demand shocks ϵ_t^N, which represent information shocks that enter her demand slowly based on the persistence parameter ϕ. If $\phi < 1$, the investor initially overreacts to shocks, but subsequently, the impact of the shock decays. For $\phi = 1$, demand is a random walk where shocks are permanent. Finally, if $\phi > 1$, a demand shock at time $t - 1$ is exacerbated further at time t. The persistence captures underreaction to information: As the investor receives a signal about fundamentals, she initially only partially adjusts her position but subsequently increasingly incorporates the information into her demand.

Market Clearing. In equilibrium, the demand of all three investors has to sum to the supply of the asset, as shown in the standard market-clearing equation (4):

$$S - D_t^N = D_t^{ST} + D_t^{LT}, \quad \forall t \tag{4}$$

Substituting demand functions (1), (2), and (3) into the market-clearing equation (4) and solving for equilibrium price changes Δp_t yields:

$$\Delta p_t = \frac{1}{\mathcal{E}_{\text{recent}}} \left(\underline{d}^{ST} - \log \left(S - D_t^{LT} - D_t^N \right) \right) \tag{5}$$

Equation (5) shows how following a demand shock from noisy demand investor N, the short-term investor is the only marginal investor willing to absorb the shock. Consequently, the price reflects her demand elasticity. Appendix A provides additional details and shows all derivations underlying results of this section.

2.2 Momentum from persistent demand shocks

Next, I show how the framework from the previous section can generate momentum. I start by emphasizing the persistence of demand shocks, representing underreaction. To illustrate the mechanism in the simplest way possible, I focus on the model without differentiating between short- and long-horizon investors. This corresponds to a flat term structure of elasticities, i.e., $\mathcal{E}_{\text{recent}} = \mathcal{E}_{\text{long-term}} = \mathcal{E}$. Then, we can aggregate the two investors into one,

$$d_t = \underline{d} - \mathcal{E} \times (p_t - p_{t-s}), \tag{6}$$

with s again capturing longer horizons of one year. To generate momentum, consider demand shocks that increase in magnitude over time, $\phi > 1$.[9] When investor N receives positive new information, she partially incorporates this into her demand, ϵ_t^N, and prices reflect the additional demand, but not enough to fully reflect the new information. Over time, the investor increasingly incorporates the news into prices; as the demand shock grows, prices increase further. This process represents underreaction to information similar to Chan, Je-

[9] To retain stationarity, one could use more complex autocorrelation patterns that lead to a build-up of shocks over short horizons but reversal over longer horizons (e.g., Lochstoer and Muir, 2022).

gadeesh, and Lakonishok (1996). In section 2.5.1, I discuss other mechanisms that generate persistent demand shocks, such as slow information diffusion of private information between many smaller investors (e.g., Hong and Stein, 1999) or delayed overreaction from belief dynamics with self-attribution bias (e.g., Daniel, Hirshleifer, and Subrahmanyam, 1998).

So what happens following at time t, following a demand shock ϵ_{t-1}^{N} that moves equilibrium returns Δp_{t-1} at $t-1$? The equilibrium follow-on return is

$$\Delta p_t = \left(\phi \frac{D_t}{D_{t-1}} - 1\right) \Delta p_{t-1} \approx (\phi - 1)\, \Delta p_{t-1}, \tag{7}$$

which is greater than Δp_{t-1} for $\phi > 1$ and sufficiently small demand shocks. Because the demand shock from time $t-1$ builds up further at t, there is additional price pressure, raising prices further: momentum.

A critical feature of this model is that while the price-sensitive investor is contrarian in her trading against prices, she does not anticipate the dynamics of the noisy investor's demand shocks. Yet even in the presence of such arbitrageurs, theories of underreaction (e.g., Hong and Stein, 1999) can still play a role in equilibrium. This is especially the case if arbitrageurs' ability to correct mispricings is subject to limits-to-arbitrage (Shleifer and Vishny, 1997) or if it is difficult to distinguish between information and noise in prices.

2.3 Momentum from the term structure of demand elasticities

Above I have shown how time-series patterns in demand shocks build up to form momentum. Next, I propose an alternative mechanism, the term structure of demand elasticities, and show how it creates momentum from investors' differential responses to price signals across horizons. For expositional purposes, I fix the persistence of demand shocks at $\phi = 1$, a random walk.

Short-term price impact. Consider an initial equilibrium perturbed by the noise trader shock, ϵ_t^N. How much does the demand shock move the equilibrium return Δp_t? To see this, first, define the effective supply $\tilde{S}$:

$$\tilde{S}_t \equiv S - D_t^{LT} - D_t^N \tag{8}$$

$\tilde{S}$ captures the effective supply after accounting for price-insensitive demand and represents the total demand the short-run investor has to absorb. We can now express the price impact of the demand shock as a function of effective supply. For example, consider a 1% shock to the supply of Tesla because Elon Musk sells shares. What happens to the price of Tesla? The answer depends on how willing the short-run investor is to absorb the shock. In the presence of a hyper-elastic ($\mathcal{E}_{\text{recent}} = \infty$) short-run arbitrageur, who responds infinitely strongly to any tiny mispricing, Tesla's price will remain anchored at its efficient level. In contrast, with inelastic demand, for example, $\mathcal{E}_{\text{recent}} = 2$ and the short-run investor owning 50% of Tesla, the short-run investor's response is $2\% \times 0.5 = 1\%$ when prices decrease by 1%, fully offsetting the size of the shock. Consequently, in equilibrium, the price of Tesla stock declines by 1%. With fewer or less elastic short-run investors, the shock is only fully absorbed when the price response grows in magnitude. More formally, define aggregate elasticities as:

$$\bar{\mathcal{E}}_{\text{recent},t} \equiv D_t^{ST} \mathcal{E}_{\text{recent}} \tag{9}$$

$$\bar{\mathcal{E}}_{\text{long-term},t} \equiv D_t^{LT} \mathcal{E}_{\text{long-term}} \tag{10}$$

Then, equilibrium condition (5) can be re-written as:

$$\Delta p_t = -\bar{\mathcal{E}}_{\text{recent},t}^{-1} \Delta \tilde{S}_t. \tag{11}$$

The price impact of a shock to the effective supply is proportional to the inverse of the aggregate recent elasticity $\bar{\mathcal{E}}_{\text{recent},t}$.[10] The more elastic the short-run investor, the steeper the demand curve she is moving along, and the more willing she becomes to absorb the demand shock at small price discounts. Thus, the less the price changes in the perturbed relative to the initial equilibrium.

Long-term price impact. Now move forward one quarter. What is the impact of a demand shock ϵ^N_{t-1} on the equilibrium price change Δp_t? Again, assume that the size of the demand shock is constant between $t-1$ and t, i.e., $\phi = 1$. From equilibrium condition (5):

$$\Delta p_t = -\frac{\bar{\mathcal{E}}_{\text{long-term},t} - \bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}}\,\Delta p_{t-1} \approx -\underbrace{\frac{\bar{\mathcal{E}}_{\text{long-term},t} - \bar{\mathcal{E}}_{\text{recent},t}}{\bar{\mathcal{E}}_{\text{recent},t}}}_{\substack{\text{term structure of}\\ \text{demand elasticities}}}\,\Delta p_{t-1} \tag{12}$$

Δp_{t-1} denotes the original price impact of the demand shock ϵ^N_{t-1} at time $t-1$. The follow-on price impact of a past shock, Δp_t, is controlled by the term structure of demand elasticities: When $\bar{\mathcal{E}}_{\text{long-term},t} = \bar{\mathcal{E}}_{\text{recent},t}$, the term structure of elasticities is flat. Due to the different investment horizons, the short-horizon investor passes the asset on to the longer-horizon investor. Still, since they are equally elastic in their aggregate responses, meaning they are equally willing to absorb shocks, they do so at the same price the short-term investor purchased the asset for. Consequently, the past demand shock has no additional impact on current prices beyond the last period's initial price impact, and $\Delta p_t = 0$. When $\bar{\mathcal{E}}_{\text{long-term},t} > \bar{\mathcal{E}}_{\text{recent},t}$, long-horizon investors are more elastic than short-horizon investors. In this case, a shock's initial price impact partially reverses. In contrast, when $\bar{\mathcal{E}}_{\text{long-term},t} < \bar{\mathcal{E}}_{\text{recent},t}$, the term structure of elasticities is downward-sloping. This is what I find to be the case in the data. Investors are less responsive to longer-term price variation, so the initial price changes must be amplified to maintain equilibrium, generating momentum.

[10] Equivalently, Gabaix and Koijen (2020) show that the price impact of aggregate flows is the inverse of their macro elasticity.

2.4 Aggregation

In reality, the distinction between short-horizon and long-horizon investors is less clear-cut than described so far. Instead, investors live on a spectrum regarding how aggressively they trade against recent versus longer-term price changes. To capture this, I leave behind the strict separation of short- and long-horizon investors and introduce a decentralized version of the model instead. It contains many investors, indexed by i, whose behavior is defined through their recent elasticity $\mathcal{E}_{\text{recent},i}$, longer-term elasticity $\mathcal{E}_{\text{long-term},i}$, and baseline demand $\underline{d}_i$:

$$d_{it} = \underline{d}_i - \mathcal{E}_{\text{recent},i} \times (p_t - p_{t-1}) - \mathcal{E}_{\text{long-term},i} \times (p_{t-1} - p_{t-s}) \tag{13}$$

Demand shocks are still the result of a separate investor with noisy demand for this section, as described in equation (3).[11] This model aggregates well; the aggregate elasticity on the stock level is equal to the demand-weighted average of elasticities across investors, similar to equations (9) and (10):

$$\bar{\mathcal{E}}_{\text{recent},t} \equiv \sum_i D_{it}\ \mathcal{E}_{\text{recent},i} \tag{14}$$

$$\bar{\mathcal{E}}_{\text{long-term},t} \equiv \sum_i D_{it}\ \mathcal{E}_{\text{long-term},i} \tag{15}$$

This enhanced model combines both channels into one equation,[12]

[11] Alternatively, I could decentralize the demand shocks to institutions as well. I do so in the data. Moreover, in the empirical measurement, I allow for arbitrary time-series patterns of demand shocks, which can differ across institutions. Therefore, my empirical findings do not require the existence of a single persistence parameter ϕ.

[12] Equation (17) corresponds to an approximation of the follow-on price change because compositional changes in ownership structure lead to time-series variation of aggregate elasticities. Equation (16),

$$\Delta p_t = \left((\phi - 1)\ \frac{\bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}} - \frac{\bar{\mathcal{E}}_{\text{long-term},t} - \bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}} \right) \Delta p_{t-1}, \tag{16}$$

$$\Delta p_t = \Bigg(\underbrace{\phi - 1}_{\substack{\text{persistent}\\ \text{demand shocks}}} - \underbrace{\frac{\bar{\mathcal{E}}_{\text{long-term}} - \bar{\mathcal{E}}_{\text{recent}}}{\bar{\mathcal{E}}_{\text{recent}}}}_{\substack{\text{term structure of}\\ \text{demand elasticities}}} \Bigg) \Delta p_{t-1}. \tag{17}$$

Equation (17) shows that the price change Δp_{t-1} from a demand shock at time $t-1$ is followed by a "momentum return" Δp_t proportional to Δp_{t-1}. When demand shocks are persistent, $\phi > 1$, a high (demand-shock-implied) return Δp_{t-1} is followed by an additional positive return Δp_t because the shock builds up further in size. This is the channel described in section 2.2. Similarly, when investors in aggregate get less responsive to the demand shock, $(\bar{\mathcal{E}}_{\text{long-term}} - \bar{\mathcal{E}}_{\text{recent}})/\bar{\mathcal{E}}_{\text{recent}} < 0$, then there again is an additional positive price change controlled by the magnitude of the downward slope of the term structure of demand elasticities.

I collect these results about the making of momentum in Proposition 1:

Proposition 1. *For price-elastic investors with demand* (13)*, noisy demand investors* (3)*, and fixed supply S, the time t follow-on price impact of a $t-1$ demand shock that initially moved prices by Δp_{t-1} is*

$$\Delta p_t \approx \left(\phi - 1 - \frac{\bar{\mathcal{E}}_{\textit{long-term}} - \bar{\mathcal{E}}_{\textit{recent}}}{\bar{\mathcal{E}}_{\textit{recent}}} \right) \Delta p_{t-1}, \tag{18}$$

where the aggregate recent elasticity $\bar{\mathcal{E}}_{\textit{recent}}$ and longer-term elasticity $\bar{\mathcal{E}}_{\textit{long-term}}$ are defined in equations (14) *and* (15)*, respectively.*

Momentum arises if:

(a) Demand shocks are persistent, $\phi > 1$.

(b) The term structure of demand elasticities is downward-sloping, $(\bar{\mathcal{E}}_{\textit{long-term}} - \bar{\mathcal{E}}_{\textit{recent}})/\bar{\mathcal{E}}_{\textit{recent}} < 0$.

exhibits the precise formulation for the follow-on price changes, incorporating a wedge between the aggregate recent elasticities as of $t-1$ and t. However, such composition-driven time-series changes are small from period to period, motivating the approximation in equation (17), which treats aggregate elasticities as locally constant.

The results from Proposition 1 follow directly from derivations in Appendix A. Proposition 1 shows that two distinct channels drive time-series patterns of price changes: the persistence of demand shocks and the aggregate term structure of demand elasticities. Momentum arises when there is a build-up in demand shocks over time, $\phi > 1$, and when the term structure of demand elasticities is downward-sloping, $(\bar{\mathcal{E}}_{\text{long-term}} - \bar{\mathcal{E}}_{\text{recent}})/\bar{\mathcal{E}}_{\text{recent}} < 0$. In contrast, mean reversion in demand shocks, $\phi < 1$, and an upward-sloping term structure of elasticities create reversals in stock returns. This paper's goal is to quantify the importance of these two channels for making momentum in equilibrium. To this end, I show how to incorporate these two channels into an asset demand system in section 3. My estimates suggest that a downward-sloping term structure of demand elasticities is the primary driver of momentum returns between 1999 and 2020.

2.5 Foundations of momentum

The framework above shows how stock momentum arises from investors' dynamic trading. It distinguishes between underreaction that manifests itself through investors' demand shocks and investors' dynamic response to prices. As I show below, many economic channels operate within these two broad categories. In practice, all of these mechanisms play some role in making momentum. By remaining agnostic about specific foundations, my empirical framework can separate the net importance of what lies at the core of creating momentum: demand shocks vis-à-vis differential responses to price changes across horizons.

First, I outline some theories of why investors would trade in a way that generates momentum through persistent demand shocks: underreaction to information, slow information diffusion, self-attribution bias, and earnings extrapolation. Second, I highlight theories that shape the term structure of demand elasticities: the evolution of arbitrage intensities across horizons, learning from prices, and the disposition effect.

2.5.1 Persistent demand shocks

Underreaction to information. Persistent demand shocks are the first mechanism through which my model can generate momentum. Some models create persistent demand shocks through underreaction to information, as in Chan, Jegadeesh, and Lakonishok (1996). Investors initially only partially react to earnings surprises. Over time, however, they increasingly incorporate the news into their demand, leading to a drift in prices. In the data, this gradual adjustment leads to persistence in demand shocks, as modeled in section 2.2, with $\phi > 1$.

Slow information diffusion. A similar example is slow information diffusion by newswatchers in Hong and Stein (1999). When some but not all investors receive private signals, then the initial total response to fundamental news is weak, underreaction. This aggregate underreaction is more pronounced if early informed investors cannot strategically front-run the demand from investors who receive the signal later. Then, as information slowly spreads, more investors respond, generating price drifts: momentum. Aggregated into one investor, this is the same mechanism as for underreaction to information. However, these channels differ in whether underreaction occurs within one investor or spread across many.

Self-attribution bias. Biased confidence dynamics, as in Daniel, Hirshleifer, and Subrahmanyam (1998) or Luo, Subrahmanyam, and Titman (2020) can also create time-series patterns in returns that resemble momentum. Investors get asymmetrically more confident when their views are validated. In particular, an investor who initially receives a positive private signal and invests will subsequently, following a positive public news release, overreact and invest too much. This overreaction stems from self-attribution bias: Investors update their positions more aggressively if observed signals align with their prior beliefs. This example of delayed overreaction leads to momentum because, more often than not, a change in demand is followed by another demand change: persistet demand shocks.

Earnings extrapolation. Extrapolation of fundamentals, for example, earnings or cash flows, can also create persistence in demand shocks. This happens in models like Barberis, Shleifer, and Vishny (1998) when investors underreact to earnings shocks because they mistakenly believe the shock to be mean-reverting, or in more recent work by De La O and Myers (2021) and Bordalo et al. (2022). However, earnings extrapolation differs from return extrapolation, which, as I show below, creates momentum through the term structure of demand elasticities. Therefore, my framework can be used to contrast the distinct roles of extrapolation based on fundamentals and prices.[13]

2.5.2 Term structure of demand elasticities

Arbitrage across horizons. Non-flat term structures of demand elasticities can arise from segmentation in arbitrage activity across different horizons. Section 2.3 outlined a model in this spirit. When two sets of arbitrageurs operate at different frequencies and differ in their aggregate willingness to absorb shocks, then prices will generally vary as the asset changes hands from being owned by the shorter-horizon to the longer-horizon arbitrageur. In particular, if short-horizon arbitrageurs are relatively more willing to absorb shocks, then the term structure of demand elasticities is downward-sloping. Specifically, long-horizon arbitrageurs might be less inclined to absorb shocks because of limits to arbitrage (e.g., Shleifer and Vishny, 1997): They might have to take on more long-run fundamental risk or might be subject to funding frictions resulting from a misalignment between the investment horizon of their assets compared to the maturity structure of their liabilities.

Learning from prices. The model of Hong and Stein (1999) features momentum traders alongside newswatchers. Momentum traders are investors who use past returns as a signal for future expected returns,[14] which is informative due to the slow information diffusion of

[13] McCarthy and Hillenbrand (2021) entertain the possibility of both return- and cash flow extrapolation and time-varying risk aversion as potential drivers for stock market fluctuations. In their estimates, they ascribe approximately equal roles to each of them.

[14] A conceptually related yet less rigorous version of learning from past prices is outright positive-feedback "trend-chasing" behavior, as in Long et al. (1990).

newswatchers' private information. In the model, momentum traders cannot post demand curves conditional on prices in the style of Grossman and Stiglitz (1980) and, therefore, only learn from past prices. Thus, the behavior of momentum traders is inelastic at short horizons, $\mathcal{E}_{\text{recent}} = 0$, and, due to the learning channel, exhibits negative elasticity to longer-term variation in prices, $\mathcal{E}_{\text{long-term}} < 0$. This combination of zero recent elasticity and negative longer-term elasticity generates a downward-sloping term structure of elasticities (i.e., $\mathcal{E}_{\text{recent}} > \mathcal{E}_{\text{long-term}}$) and hence, momentum in stock returns.[15][16] More generally, learning from prices leads to more inelastic demand (Haddad, Huebner, and Loualiche, 2022), especially for uninformed investors who cannot distinguish between information and noise in prices (Davis, Kargar, and Li, 2022).[17][18] To the extent that learning from prices is instantaneous and long-lasting, it shifts elasticities at all horizons up or down but does not create differential behavior in relative terms. However, there are many examples of deviations from these assumptions. First, Davis (2021) argues that across many canonical portfolio choice models (e.g., Brandt, Santa-Clara, and Valkanov, 2009), investors learn about expected returns from past returns but do not post demand curves that enable learning from current equilibrium prices (e.g., Grossman and Stiglitz, 1980; Veldkamp, 2011). Second, past returns can enter belief formation through weights that are not constant across time. Richer term structures of elasticities represented through more than two elasticities could model such patterns in belief-formation weights. Non-constant weights in belief formation can occur rationally when investors learn about moving targets (e.g., Collin-Dufresne, Johannes, and Lochstoer, 2016). Alternatively, it can reflect a wedge between subjective and objective expectation formation, for example, when investors' lived experiences decay slowly (Malmendier and Nagel, 2011, 2016; Nagel and Xu, 2022) or as the consequence of investors extrapolating past returns (Bar-

[15] More precisely, momentum traders exacerbate the price drift caused by slow information diffusion. Relatedly, Hong, Lim, and Stein (2000) show that momentum strategies work better for stocks with slow information diffusion, as proxied by less analyst coverage and smaller firm size.

[16] Lou and Polk (2021) show that the larger the momentum crowd, that is, the more momentum traders there are, the more prices overshoot fundamentals and revert subsequently.

[17] See Adam and Nagel (2022) for a review of the role of expectation formation in asset pricing.

[18] Consistent with this idea, the pass-through from exogenous variation in prices to investors' expected returns (Charles, Frydman, and Kilic, 2022; Chaudhry, 2022) and portfolios (Giglio et al., 2021) is weak.

beris and Shleifer, 2003; Greenwood and Shleifer, 2014; Barberis et al., 2015, 2018; Cassella and Gulen, 2018).[19] Notice how in contrast to earnings extrapolation, which in the above framework works through persistent demand shocks, return extrapolation affects the term structure of demand elasticities.

Disposition effect. As a final example, Grinblatt and Han (2005) show that the disposition effect, investors' tendency to sell winners too early and losers too late, can generate momentum in stock prices. Consider a setting in which investors' demand has a rational component based on deviations of equilibrium prices from fundamental values but also features deviations of equilibrium prices from some perceived reference prices, their cost bases. This corresponds to a demand function with two elasticities. As long as the reference price corresponds to past stock prices, it is equivalent to two elasticities for different time horizons. To align the time horizons with momentum frequencies, consider the stock price from one quarter ago as the reference price. Then investors overreact to recent price changes; that is, they are more elastic to variation in prices over the most recent quarter, $\mathcal{E}_{\text{recent}} > \mathcal{E}_{\text{long-term}}$, which corresponds to a downward-sloping term structure of elasticities and generates momentum.

3 Estimating Dynamic Trading

In this section, I estimate the two channels that create momentum in equilibrium: the evolution of demand shocks and the term structure of demand elasticities. I start by putting forth a demand system in the style of Koijen and Yogo (2019) that accounts for equilibrium and incorporates both demand shocks and the term structure of demand elasticities. Then, I introduce a novel identification strategy for demand estimation in the presence of dynamic trading and implement it for all institutional investors in the U.S. stock market between 1999

[19] As emphasized by Da, Huang, and Jin (2021), extrapolation models generate reversals because the impact of past shocks decays over time. However, they differ in terms of their ability to generate momentum based on whether investors' response to past returns is hump-shaped across horizons. This occurs when investors do not immediately incorporate returns into belief formation. Once they do, their impact starts to decay.

and 2020.

3.1 Quantitative model

Investor demand. Unlike in the model from section 2, investors choose portfolios of stocks. Koijen and Yogo (2019) show that a logit of portfolio weights is a good way of modeling portfolio choice, as it ensures that portfolio weights for each investor sum to 1 and allows for substitution across assets.[20] I follow this approach. In particular, I use a log-linear specification to model portfolio weights relative to an outside asset 0, $\log\left(w_{it}(n)/w_{it}(0)\right)$, where $w_{it}(n)$ indexes the investor i's portfolio weight in stock n at time t. The resulting portfolio demand is

$$\underbrace{\log \frac{w_{it}(n)}{w_{it}(0)}}_{\text{demand}} = \underbrace{(1-\mathcal{E}_{\text{recent},i})\ \Delta p_t(n)}_{\text{contemporaneous elasticity}} + \underbrace{(1-\mathcal{E}_{\text{long-term},i})\ \left(\sum_{s=1}^{3} \Delta p_{t-s}(n)\right)}_{\text{dynamic elasticity}} + \underbrace{\underline{d}_{0it} + \underline{d}_{1i}' X_t(n)}_{\text{characteristics demand}} + \underbrace{\epsilon_{it}(n)}_{\text{latent demand}}. \tag{19}$$

The first two components of the demand system capture price-elastic demand: when the price of an asset rises, investors' demand for it decreases. The larger the elasticities, the more aggressive is the investor in trading against prices. In contrast to previous studies, I allow investors to respond differentially to recent and longer-term variation in prices, which is captured through two separate parameters, $\mathcal{E}_{\text{recent},i}$ and $\mathcal{E}_{\text{long-term},i}$. When an investor has $\mathcal{E}_{\text{recent},i} > \mathcal{E}_{\text{long-term},i}$, I say that this investor has a downward-sloping term structure of elasticities. In order to align my elasticity estimates with the time horizons in the momentum literature (Jegadeesh and Titman, 1993), I separate the price change over the last year into the most recent quarter and the three preceding quarters. That is, I model the demand as of December 31 as a log-linear function of the return between October and December 31 (the contemporaneous price-elastic demand), and the return between December 31 of the

[20] Koijen and Yogo (2020) allow for more flexible substitution patterns across countries and asset classes.

previous year and October 31 (the longer-term price-elastic demand). In the estimation, I impose downward-sloping demand curves for the contemporaneous elasticity, $\mathcal{E}_{\text{recent},i} \geq 0$, which is necessary for the decomposition of section 4 (Koijen and Yogo, 2019). However, I do allow negative longer-term elasticities to capture learning from past prices, or trend-chasing more generally, as is the case for momentum traders in Hong and Stein (1999).[21]

The third component of the demand function is $\underline{d}_{0it} + \underline{d}_{1i}' X_t(n)$ and captures investor-specific functions of common stock characteristics. I include book equity, profitability, investment, and dividend yield. These characteristics can be used by investors to form beliefs about firm fundamentals and expected returns. Finally, latent demand captures unobserved demand shocks. Such shocks may correspond to private information, but could also capture investor tastes or noise trading.

My model does not, strictly speaking, nest the constant elasticity model of Koijen and Yogo (2019), because I do not allow for flexible long-term elasticities beyond the horizon of one year. Instead, I assume that long-term elasticities are 1 across investors.[22]

Investor assets. While the assets-under-management process is less important for the estimation of investor portfolio demand, it does play a role in counterfactuals: If the return to an asset an institution holds had been different, the evolution of its asset dynamics would have changed as well. Therefore, I partially endogenize the asset dynamics of institutions. That is, I separate out the portions of asset dynamics that are endogenous through portfolio returns from a flow component, which I consider invariant to the equilibrium. This is unlike previous papers in the demand-system literature, which treat the evolution of an institution's

[21] One might be concerned that downward-sloping term structure of elasticities estimates might simply be the result of allowing negative elasticities with respect to long-term but not recent returns. If that were the case, I should find strongly negative elasticity term structures among the initially inelastic investors. This is counterfactual to the estimation results depicted in Figure 1, and therefore unlikely to pose an issue.

[22] In an alternative specification, I consider subtracting prices from the left-hand side of equation (19) and omit the ones as parts of the contemporaneous- and dynamic elasticity terms. This would correspond to setting investors' unmodeled long-term elasticities to zero. However, it would generate strong momentum at long horizons, which is counterfactual to long-term overreaction and reversal (e.g., De Bondt and Thaler, 1985). Therefore, I use a long-term target elasticity of 1, which is slightly higher than the average elasticities estimated in Koijen and Yogo (2019).

assets under management as exogenous.

$$A_{it} = A_{it-1}\left(1 + f_{it} + w_{it-1}(n)'\Delta p_t(n)\right), \quad \forall i. \tag{20}$$

The assets under management A_{it} of institutions in equation (20) are functions of past assets A_{it-1}, flows f_{it} and equilibrium portfolio returns $w_{it-1}(n)'\Delta p_t(n)$.[23]

Equilibrium returns. Equilibrium returns are determined as market-clearing returns, solving the equilibrium of individual demands. Normalizing the number of shares to 1, the market-clearing equation for the log equilibrium return is

$$\Delta p_t(n) = p_t(n) - p_{t-1}(n) = \log\left(\frac{\sum_i A_{it} w_{it}(n)}{\sum_i A_{it-1} w_{it-1}(n)}\right), \quad \forall n, \tag{21}$$

where the portfolio weight $w_{it}(n)$, and thus the right-hand-side of equation (21), is decreasing in the return $\Delta p_t(n)$.[24] This guarantees the existence of equilibrium for the decomposition in Section 4.1.[25]

Momentum from dynamic investor trading. As I demonstrated in section 2, the model can generate momentum from two dimensions of investor trading. First, momentum is generated from investors with a downward-sloping term structure of elasticities. These are investors who react to prices in a more aggressively contrarian way at short horizons, but

[23] Unlike for mutual funds, exact flows and returns for institutions are not readily available in the data. I manually separate them by making an assumption about the timing of portfolio changes between quarter-end cutoff dates: I assume that institutions keep their quarter-end holdings until just before the next quarter-end. Under this assumption, I can separate out an institution's portfolio return $w'_{it-1}\Delta p_t(n)$ and reverse-engineer inflows as $f_{it} \equiv (A_{it} - A_{it-1})/A_{it-1} - w'_{it-1}\Delta p_t(n)$.

[24] Technically, there is also the wealth effect from equation (20). As I show in Appendix B.3, this effect can, in principle, generate negative elasticities for passive investors with concentrated portfolios. Practically, however, I do not find this to be of issue. In particular, in counterfactual exercises my numerical algorithm converges to an equilibrium within few iterations.

[25] For uniqueness of the equilibrium, there needs to be at least one non-passive investor with $\mathcal{E}_{\text{recent},i} > 0$ (Haddad, Huebner, and Loualiche, 2022) in the stock. The condition is satisfied for every stock at each time.

subsequently become less aggressive. In the data, I find strong support for this mechanism, which could reflect the dynamics of arbitrage, price-chasing behavior, or learning from past prices.

Second, there is the evolution of demand shocks, $\epsilon_{it}(n)$. This is the component of the demand system that captures many theories of underreaction. For example, when an investor receives a private signal, she will incorporate it into her latent demand. But if initially, she does not fully incorporate the information into her demand, then there will be persistence in her demand shocks, which is underreaction. Latent demand will capture both the dynamics of underreaction within the same investor across time and underreaction "in aggregate", which occurs when a demand shock of some investor predicts future shocks of others. That is, underreaction can occur within the same investor, but it can also occur when some investor has early access to information, and information diffusion is slow (Hong and Stein, 1999). Either way, it generates persistence in aggregate latent demand.

3.2 Data

I follow Koijen and Yogo (2019) and Haddad, Huebner, and Loualiche (2022) in obtaining stock-level data and data on portfolio holdings for the U.S. stock market. Data on stock prices, returns, dividends, and shares outstanding are from CRSP, and book equity, profitability, and investment are from COMPUSTAT.

In addition, I source data on institutional investors' portfolio holdings between Q4 1999 and Q4 2020 from regulatory 13F filings available on the SEC EDGAR website using the method of Backus, Conlon, and Sinkinson (2019, 2020). Institutions with at least $100mn in assets under management are required to file quarterly reports of their entire stock positions to the SEC, which sums to a total coverage of about 80% of total U.S. stock market capitalization. I follow Koijen and Yogo (2019) in grouping the remainder in an investor that I label the household sector.[26]

[26] I use the term "household sector" in a slight abuse of language, as it captures direct household holdings alongside, for example, holdings by small institutions below the reporting threshold.

Finally, I obtain mutual fund data from the CRSP Survivor-Bias-Free US Mutual Fund Database. It contains information on mutual fund flows, returns, and holdings,[27] all used to construct the instrument for returns: flow-induced trading (Lou, 2012).

3.3 Identification

3.3.1 Identification problems

By substituting portfolio demand (equation (19)) into market-clearing (equation (21)), one can immediately see that latent demand affects equilibrium returns: positive demand shocks put upwards-price pressure on prices. Moreover, demand shocks may be correlated across investors. Both lead to mechanical correlation between returns and latent demand, i.e. $cov\left(\epsilon_{it}(n), \Delta p_t(n)\right) \neq 0$, and therefore introduce a bias in estimating the real-time elasticity $\mathcal{E}_{\text{recent},i}$, which is the investor's response to the return $\Delta p_t(n)$, via OLS. This is the standard simultaneity issue common to any setting of demand estimation. Below, I introduce an instrument that allows me to disentangle an investor's response to contemporaneous returns from their demand shocks.

But first, there is also a dynamic simultaneity issue specific to my setting. To see this, think of an investor with an underreaction type of demand shock. For example, at time $t-1$ an investor, Elon, receives a positive private signal and buys some shares of Tesla. At time t, he buys even more.[28] In such a setting, it is difficult to disentangle investors' dynamic responses to the shock from the dynamics of the shock itself. Investor demand correlates with longer-term price changes, but is that because of the response we want to identify — how investors react to long-term returns — or because of investors reacting to Elon's additional buying of Tesla stock at time t? More formally, consider the moment condition under a valid instrument for returns, with $\widehat{\Delta p_t}(n)$ denoting instrumented returns:

[27] Like Dou, Kogan, and Wu (2020), I use the CRSP mutual fund holdings data as of Q3 2008, but the Thomson Reuters Mutual Fund Holdings Data prior to that date.

[28] Slow trading by insiders can be optimal in models in which insiders try to conceal their private information (e.g., Kyle, 1985).

$$\mathbf{E}_i\left[\epsilon_{it}(n)|\mathbf{X}_t(n), \widehat{\Delta p}_t(n), \sum_{s=1}^{3}\Delta p_{t-s}(n)\right] = 0, \quad \forall i \tag{22}$$

This moment condition requires that latent demand $\epsilon_{it}(n)$ would have to be uncorrelated with the past returns $\Delta p_{t-1}, \Delta p_{t-2}$, and Δp_{t-3}. However, past returns are themselves equilibrium objects and have to satisfy the market clearing equations at time $t-1$, $t-2$, and $t-3$, respectively. By the exact same argument as for the standard simultaneity issue, $cov\left(\epsilon_{it-1}(n), \Delta p_{t-1}(n)\right) \neq 0$. This implies that the only way that $\epsilon_{it}(n)$ can be orthogonal to $\Delta p_{t-1}(n)$ is if latent demand itself is uncorrelated across time, i.e. $cov(\epsilon_{it}(n), \epsilon_{jt}(n)) = 0, \forall j$. However, the assumption of uncorrelated demand shocks across time is rejected both by the data and conceptually, as it rules out any momentum- or reversal generating persistence of demand shocks.

The dynamic simultaneity issue reflects a combination of persistent demand shocks and classic simultaneity issues. In order to solve it and identify investors' response to longer-term variation in returns, $\mathcal{E}_{\text{long-term},i}$, I proceed in a way that is analogous to solving the classic simultaneity problem: I isolate exogenous variation in longer-term price changes through an instrument orthogonal to $\epsilon_{it}(n)$. Assuming a valid instrument for longer-term returns, the moment condition then weakens to

$$\mathbf{E}_i\left[\epsilon_{it}(n)|\mathbf{X}_t(n), \widehat{\Delta p}_t(n), \sum_{s=1}^{3}\widehat{\Delta p}_{t-s}(n)\right] = 0, \quad \forall i. \tag{23}$$

3.3.2 Instruments for recent and long-term returns

I proceed by introducing instruments for recent and longer-term returns. My instrument for recent returns is based on mutual fund flow-induced trading from Lou (2012).[29] The idea

[29] Flow-induced trading can be viewed as a generalization of mutual fund fire sales induced flows as in Coval and Stafford (2007).

of this instrument is that when mutual funds face redemptions, they are forced to partially liquidate their holdings. Assuming that funds sell proportional to their past holdings, a mutual fund's flows will generate cross-sectional variation in price pressure proportional to the fund's holdings.[30] The instrument then aggregates this flow-induced price pressure across funds on the stock level.

$$FIT_t(n) \equiv \sum_j \frac{A_{jt-1} w_{jt-1}(n)}{P_{t-1}(n)} f_{jt} = \sum_j o_{jt-1}(n) f_{jt} \tag{24}$$

Equation (24) shows the definition of flow-induced trading more formally. Subscript j captures mutual funds, which is in contrast to before when variables were defined on the institution level more broadly. $P_{t-1}(n)$ captures the $t-1$ market capitalization of stock n, f_{jt} are the net inflows fund j received between $t-1$ and t, and o_{jt-1} captures the share fund j holds of stock n at time $t-1$.

Appendix section B.2 derives equation (24) by starting at the market clearing equation for returns (21), and making three adjustments to avoid sources of endogeneity: (i) replacing endogenous assets A_{jt} by $A_{jt-1}(1+f_{jt})$, (ii) replacing current portfolio weights $w_{jt}(n)$ by past weights $w_{jt-1}(n)$, and (iii) filtering from the set of all investors to mutual funds only. The latter is due to the availability of mutual-fund flow data, which is not the case for all financial institutions more generally. On the flip side, the variation that the instrument does use comes from mutual fund flows and past mutual-fund ownership: Stocks that last period were owned by mutual funds that subsequently received a lot of inflows have high flow-induced trading.

The key identification assumption is that mutual-fund flows are uncorrelated with latent demand. Yet, there is robust evidence that mutual fund flows follow past fund performance. For example, retail investors might use past fund returns to learn about fund manager skill.[31]

[30] Mutual funds could smooth in- and outflows through cash holdings, or not scale holdings up or down proportionally. Lou (2012) shows that indeed the pass-through of redemptions to proportional selling is close to 1 for 1, but somewhat lower for inflows, where only about 60 to 80 cents of each inflow dollar are used to scale up existing holdings.

[31] Early empirical evidence of the flow-fund relationship include Ippolito (1992), Chevalier and Ellison

Could the flow-performance relation potentially induce correlation between fund flows and latent demand? Yes, because as I have argued above in the context of the dynamic simultaneity issue, persistent demand shocks can lead to correlation between between $\epsilon_{jt}(n)$ and past returns. In order to overcome the identification problem posed by the fund-performance relation, I orthogonalize mutual fund flows with respect to past fund performance and past fund flows. That is, I regress quarterly mutual fund flows on the fund flows and fund performance of the four proceeding quarters, and extract orthogonalized flows $\tilde{f}_{jt}$. The regression results are shown in Appendix Table IA.1.[32] Controlling for past fund performance and past flows allows me to isolate exogenous components of mutual fund flows. I then construct orthogonalized flow-induced trading, $\widetilde{FIT}_t(n)$, analogously to before:

$$\widetilde{FIT}_{t-1\to t}(n) \equiv \sum_j o_{jt-1}(n)\tilde{f}_{jt} \tag{25}$$

In addition to the instrument for contemporaneous returns, I also require an instrument for longer-term returns, $\sum_{s=1}^{3} \Delta p_{t-s}(n)$. I proceed analogously to above, and define the instrument for longer-term returns as

$$\widetilde{FIT}_{t-4\to t-1}(n) \equiv \sum_j o_{jt-4}(n)\left(\tilde{f}_{jt-3} + \tilde{f}_{jt-2} + \tilde{f}_{jt-1}\right). \tag{26}$$

Relevance condition. Table 1 shows the results from first-stage regressions of recent and longer-term returns onto recent and longer-term flow-induced trading. In particular, columns 1 and 2 show first-stage results for current returns, while columns 3 and 4 focus on momentum-frequency returns. Columns 1 and 3 use raw FIT as proposed by Lou (2012).

(1997), and Sirri and Tufano (1998). Berk and Green (2004) is an example of a rational model that incorporate retail investors learning about mutual fund skill.

[32] I use specification 2 from Appendix Table IA.1, which introduces time-fixed effects to control for time-series variation in aggregate flows.

Table 1. Relevance conditions for the recent and longer-term return instruments.

	Return $p_t - p_{t-1}$		Past Return $p_{t-1} - p_{t-4}$	
	(1)	(2)	(3)	(4)
$FIT_{t-1\to t}(n)$	1.399***		1.430***	
	(0.233)		(0.310)	
$FIT_{t-4\to t-1}(n)$	-0.354***		1.396***	
	(0.095)		(0.207)	
Orthogonalized $\widetilde{FIT}_{t\to t-1}(n)$		1.576***		0.338
		(0.285)		(0.396)
Orthogonalized $\widetilde{FIT}_{t-4\to t-1}(n)$		-0.280*		2.298***
		(0.130)		(0.197)
Date Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
N	257,941	257,941	257,941	257,941
R^2	0.216	0.216	0.137	0.136
F	22.000	24.169	49.638	56.982
F-test p value	0.000	0.000	0.000	0.000

Table 1 reports first-stage regressions of returns over the most recent quarter, $p_t - p_{t-1}$, and the three preceding quarters, $p_{t-1} - p_{t-4}$, onto flow-induced trading instruments between 1999 and 2020. Specifications (1) and (3) use flow-induced trading, as defined in equation (24), based on Lou (2012). Specifications (2) and (4) employ the enhanced instruments, as defined in equations and (25) and (26). That is, they are based on mutual-fund flows orthogonalized with respect to past fund flows and fund returns. All specifications use date-fixed effects and control for cross-sectionally de-meaned and standardized stock characteristics: log book equity, profitability, investment, and dividend yield. Standard errors are 2-way clustered by date and stock.

In contrast, columns 2 and 4 use my orthogonalized flow-induced trading measures, which constitute the basis for my empirical findings. All regressions include time-fixed effects and controls for profitability, investment, book equity, and dividend yield.

Across all regressions, the F statistic is above 10, and instruments are strongly statistically significant based on standard errors that are two-way clustered by date and stock. Coefficients on returns at the same horizon range between 1.4 to 2.3. A coefficient of 1 would be interpreted as a flow-induced inflow of 1% to a stock predicting a 1% return of the same stock.

The orthogonalized instruments in columns 2 and 4 cleanly separate instrumenting for longer-term and recent returns. While this does not constitute a formal test of the exclusion restrictions, it is nevertheless reinsuring because it suggests the instrument is only correlated with returns at the same horizon. Otherwise, one could be worried that the instrument does not affect demand based on the price-pressure channel described above but because it is correlated with expected returns going forward. However, that does not appear to be the case in terms of longer-term returns.

I use the approach of Two-Sample Two-Stage Least Squares (Arellano and Meghir, 1992; Angrist and Krueger, 1992), meaning I estimate the first- and the second stage from different samples. This constitutes a deviation from Koijen and Yogo (2019) and Koijen, Richmond, and Yogo (2020), who estimate both within an investor's investment universe, defined as stocks the investor has held within the past three years.[33] However, investors might not only use stocks they held in the past in their formation of expected returns, which is connected to the first stage. Consequently, I relax this assumption and allow investors to learn from the entire cross-section of stocks, irrespective of which stocks they hold or are in their investment universe. Yet I do follow Koijen and Yogo (2019) in estimating the second stage within an investor's investment universe, as for many investors, the portfolio weights in most stocks are zero.

My approach has an additional, more practical advantage. Exogenous yet relevant instruments for returns that are readily available for all stocks at all times are rare, especially for investors who hold relatively few stocks and have short time series of data available to begin with. Using the full panel of stocks and time in the first stage, I can satisfy relevance conditions without excluding or grouping investors with few observations.

[33] This approach has a potential identification issue coming from investors' investment universes being potentially larger than identified from past holdings, with investors endogenously not holding certain stocks. Under such a model, the stocks with low expected returns within the investment universe will be omitted from the formation of expected returns in the first stage.

3.4 Estimates

I estimate the model between 1999Q4 and 2020Q4 for each institution using a panel approach.[34] That is, for each institution I obtain one estimate for the recent elasticity $\mathcal{E}_{\text{recent},i}$ and longer-term elasticity $\mathcal{E}_{\text{long-term},i}$.

Figure 1 visualizes my estimates for recent and long-term elasticities through a scatterplot. Each point represents an institution with recent elasticity $\mathcal{E}_{\text{recent},i}$ on the x-axis and longer-term elasticity $\mathcal{E}_{\text{long-term},i}$ on the y-axis. The black dashed line has intercept zero and slope 1, meaning that any institution below the line has $\mathcal{E}_{\text{long-term},i} < \mathcal{E}_{\text{recent},i}$, or a downward-sloping term structure of elasticities. The thick blue line is a fitted trend line based on a cubic regression.

As the smoothed blue line indicates, for institutions with recent elasticity $\mathcal{E}_{\text{recent},i}$ below 2.5, recent and longer-term elasticity are on average the same. Consequently, for low elasticity institutions, the term structure of elasticities is approximately flat. The types of institutions in this corner of the figure would include large institutional asset managers such as Fidelity who have elasticities close to zero across horizons (Haddad, Huebner, and Loualiche, 2022). Notably, the residual household sector also falls into this sector.

To the right of a real-time elasticity of 2.5, the trend line diverges from slope 1. Such institutions, on average, have downward-sloping elasticity term structures with $\mathcal{E}_{\text{long-term},i} < \mathcal{E}_{\text{recent},i}$. This class of institutions broadly captures arbitrageurs, who initially are willing to respond very elastically to shocks. Subsequently, however, they are less willing to do so, as captured by their downward-sloping elasticity term structures. In section 4.2, I show that this behavior is a major driver of momentum in the cross-section of stocks.

As of Q1 2016, about a third of institutions have elasticities above 2.5, representing about 48% of assets under management, and including large asset management firms such as Citadel LLC or Berkshire Hathaway. Another example of institutions in this area is AQR

[34] I follow Koijen, Richmond, and Yogo (2020) and a robustness specification in Haddad, Huebner, and Loualiche (2022) in using a panel approach. In contrast, other demand-system studies (e.g., Koijen and Yogo, 2019) estimate cross-sectionally and produce separate estimates at each point in time.

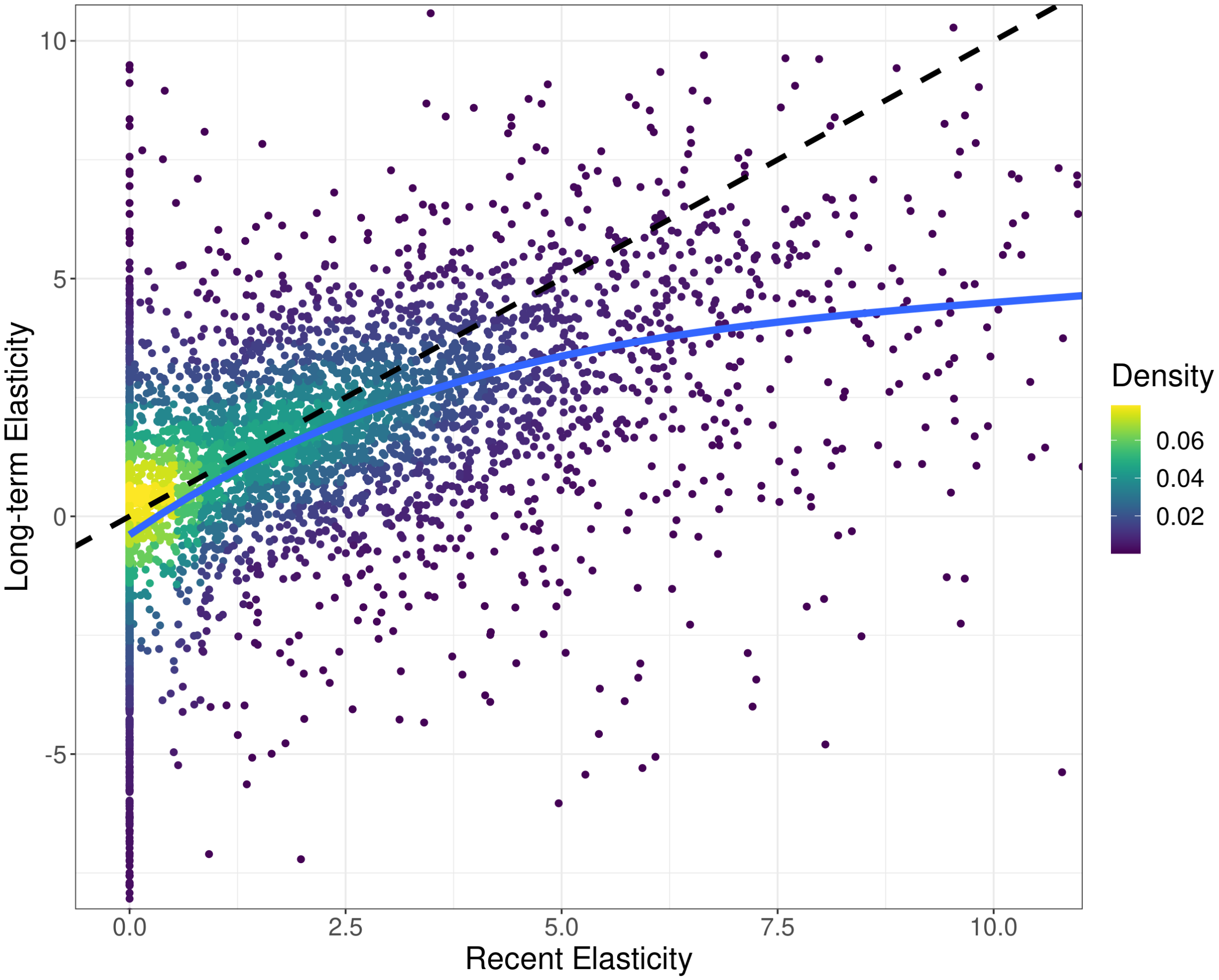

Figure 1. Estimates for elasticities $\mathcal{E}_{\textbf{recent},i}$ and $\mathcal{E}_{\textbf{long-term},i}$
Figure 1 shows a scatterplot of elasticity estimates for elasticities with respect to price changes over the past quarter, $\mathcal{E}_{\text{recent},i}$, on the x-axis, and variation over the three preceding quarters, $\mathcal{E}_{\text{long-term},i}$, on the y-axis. Each dot represents one institutional investor in the sample. The solid blue line is a fitted trend line based on cubic regression, and the black dashed line represents flat term structures of elasticities, $\mathcal{E}_{\text{long-term},i} = \mathcal{E}_{\text{recent},i}$. Dots below the dashed line are institutions with downward-sloping term structures of elasticities. The estimation equation is equation (19).

Capital Management, one of the strongest proponents of factor investing, and is particularly active in the spaces of value- and momentum investing.[35] As a value investor, AQR seeks to overweight cheap and underweight expensive stocks: when a stock becomes cheap, AQR wants to hold more of it. Such a contrarian strategy can be expressed through a high price

[35] `https://www.aqr.com/Insights/Systematic-Investing`

Table 2. Summary statistics for the term structure of elasticities $\mathcal{E}_{\textbf{long-term}} - \mathcal{E}_{\textbf{recent}}$

	$\mathcal{E}_{\text{recent}}$	$\mathcal{E}_{\text{long-term}} - \mathcal{E}_{\text{recent}}$
Average	1.93	-0.55
Standard Deviation	2.35	2.12
Quantile 10%	0.00	-2.86
Quantile 25%	0.00	-1.40
Quantile 33%	0.30	-0.97
Median	1.31	-0.33
Quantile 67%	2.29	0.25
Quantile 75%	2.92	0.59
Quantile 90%	4.77	1.52

Table 2 reports summary statistics for the cross-institution distribution of recent elasticities, $\mathcal{E}_{\text{recent}}$, and the term-structure of elasticities, $\mathcal{E}_{\text{long-term}} - \mathcal{E}_{\text{recent}}$, based on estimates of the model described in equation (19) using data between 1999 and 2020.

elasticity of demand. Indeed, AQR has a real-time elasticity $\mathcal{E}_{\text{recent}}$ of 4.15, which in Q1 2016 corresponds to the 85^{th} percentile in the cross-section of institutions. On the flip side, AQR is also a strong proponent of momentum investing, which can be expressed through longer-term elasticities lower than recent elasticities. Consistent with this, AQR's difference between recent and long-term elasticity, $\mathcal{E}_{\text{long-term}} - \mathcal{E}_{\text{recent}}$, is about -1.25, which corresponds to the 25^{th} percentile across institutions.

Table 2 shows time-series averages of cross-investor summary statistics for real-time elasticities $\mathcal{E}_{\text{recent}}$ and differences between real-time and long-term elasticity, $\mathcal{E}_{\text{long-term}} - \mathcal{E}_{\text{recent}}$. Time-series variation of these measures is purely driven by composition effects, as I estimate one real-time and one longer-term elasticity for each investor, similar to Koijen, Richmond, and Yogo (2020).

Median and average real-time elasticities are about $1.3-1.9$, which is substantially higher than constant elasticity estimates from previous asset-demand systems, by a factor of about 3 (Koijen and Yogo, 2019; Gabaix and Koijen, 2020; Haddad, Huebner, and Loualiche, 2022),[36] but in line with some estimates from other research designs (e.g., Pavlova and Sikorskaya, 2022).[37] Yet all of these are at least three orders of magnitude below the elasticity implied from a standard frictionless model (Petajisto, 2009).[38]

Around 25% of investors have recent elasticity $\mathcal{E}_{\text{recent}}$ equal to zero.[39] On the other side, about 10% of investors have elasticities above 5.

The difference between recent and long-term elasticity captures the term structure of elasticities and can drive momentum. When investors are initially willing to trade against a shock but subsequently leave, the initial price impact of the shock has to increase in equilibrium. This channel would be parametrized through $\mathcal{E}_{\text{long-term}} - \mathcal{E}_{\text{recent}} < 0$, meaning that investors' initial response to a shock $\mathcal{E}_{\text{recent}}$ is larger than they response $\mathcal{E}_{\text{long-term}}$ to a past shock, which corresponds to a downward-sloping term structure of elasticities. The cross-sectional average and median differences across investors are about -0.33 to -0.5, meaning that investors' responses to past shocks are typically about 25% weaker than their immediate responses.

There is substantial heterogeneity across investors in how they respond to prices dynamically. On the one hand, there are investors with strongly decreasing term structure of elasticities. For example, the fraction of investors whose elasticity with respect to longer-term variation in prices is lower than to recent variation in prices, by at least 1, is about 33%. On the other hand, there are also about 15% of investors whose long-term elasticity is higher than their recent elasticity by at least 1.

[36] A notable exception to this is van der Beck (2022), who also uses flow-based identification to find similar magnitudes for elasticities.

[37] See Gabaix and Koijen (2020) for a detailed summary of elasticity estimates from the literature.

[38] Davis, Kargar, and Li (2022) argue that information frictions among uninformed investors can rationalize inelastic demand curves.

[39] The estimation procedure imposes that real-time elasticities $\mathcal{E}_{\text{recent}}$ have to be non-negative, as otherwise the existence of equilibrium in the counterfactuals of section 4.1 would not be guaranteed (Koijen and Yogo, 2019).

3.4.1 Estimates aggregated on stock level

Above I have argued that there is a large degree of heterogeneity in investors' term structure of demand elasticities. I use this variation in section 4.2, combined with variation from ownership structure across stocks. Individual investor heterogeneity aggregates up to stock-level heterogeneity, based on variation in ownership across stocks. This source of variation allows me to predict where momentum should be the strongest.

More precisely, I aggregate investor-level recent and longer-term elasticities into an aggregate stock-level elasticity term structure, equivalent to equations (14) and (15):

$$\bar{\mathcal{E}}_{\text{recent},t}(n) \equiv \sum_i o_{it}(n)\mathcal{E}_{\text{recent},i} \tag{27}$$

$$\bar{\mathcal{E}}_{\text{long-term},t}(n) \equiv \sum_i o_{it}(n), \mathcal{E}_{\text{long-term},i} \tag{28}$$

$$\eta_t(n) \equiv \frac{\bar{\mathcal{E}}_{\text{long-term},t} - \bar{\mathcal{E}}_{\text{recent},t}}{\bar{\mathcal{E}}_{\text{recent},t}} \tag{29}$$

Here the ownership share $o_{it}(n)$ captures the proportion of shares investor i holds of stock n at time t, such that $\bar{\mathcal{E}}_{\text{recent},t}(n)$ and $\bar{\mathcal{E}}_{\text{long-term},t}(n)$ are the ownership-weighted average real-time and longer-term elasticities for stock n at time t, respectively. The variable $\eta_t(n)$ then captures the aggregate term structure of elasticities in the stock, like in section 2.

Table 3 provides the results of panel regressions of aggregate real-time elasticities $\bar{\mathcal{E}}_{\text{recent},t}(n)$ and term-structures of elasticities $\eta_t(n)$ onto stock characteristics.

Stocks with a log market capitalization of one standard deviation above average have recent elasticities $\bar{\mathcal{E}}_{\text{recent}}$ that are 0.7 standard deviations above average. This is consistent with the idea that large stocks are more liquid, a common result in the asset-demand system literature as liquidity and elasticities are conceptually related (Koijen and Yogo, 2019; Haddad, Huebner, and Loualiche, 2022). Beyond size, elastic stocks tend to be profitable and have a low dividend yield.

Table 3. Aggregate term structures of elasticities and stock characteristics

	Recent Elasticity	Elasticity Term Structure
	(1)	(2)
Log Market Capitalization	0.711***	0.269***
	(0.018)	(0.020)
Log Book Equity	-0.120***	0.044*
	(0.017)	(0.020)
Profitability	0.121***	0.163***
	(0.008)	(0.010)
Investment	0.007	-0.015**
	(0.005)	(0.005)
Dividend Yield	-0.210***	-0.111***
	(0.012)	(0.010)
Date Fixed Effects	Yes	Yes
N	257,941	257,941
R^2	0.395	0.139

Table 3 reports coefficient estimates from panel regression of elasticities onto stock characteristics: log market capitalization, log book equity, profitability, investment, dividend yield. The dependent variables are the aggregate recent elasticity, $\bar{\mathcal{E}}_{\text{recent},t}(n)$, in the first column, and the aggregate term structure of elasticities, $\eta_t(n)$, in the second column. All variables, including elasticities, are cross-sectionally demeaned and standardized at each date. Both specifications include date-fixed effects. The sample period is between 1999 and 2000. Standard errors are 2-way clustered by date and stock.

Similarly, stocks with one standard deviation higher log market capitalization tend to have about 0.25 standard deviations more upward-sloping elasticity term structures. And again, profitable stocks tend to have increasing elasticity term structures, while high dividend-yield stocks tend to have more decreasing term structures.

4 Implications for the making of momentum

4.1 Decomposing momentum returns

In this section, I provide a positive account of momentum returns between 1999 and 2020. I decompose momentum into how much of it results from the persistence of demand shocks and how much is due to the term structure of demand elasticities. This is where the asset demand system provides unique insights due to its ability to account for equilibrium. In the demand system, observed prices at each point in time are the equilibrium of the individual behavior of all investors. In other words, by taking all components of the demand system — stock characteristics (excluding recent equilibrium returns), parameter estimates from the demand system, including residual latent demand $\epsilon_{it}(n)$, and investor assets — one can reconstruct the market clearing equilibrium stock price, or equivalently, equilibrium return. Next, I evaluate each component's role in the demand system by tracing their evolution from time $t-1$ to time t, at each step solving for the counterfactual market clearing price, and combine them into counterfactual momentum portfolio returns based on classic momentum sorts.

More formally, I follow Koijen and Yogo (2019) in defining a function $\mathbf{g}$ that maps time-invariant demand system estimates $\theta \equiv \{\mathcal{E}_{\text{recent},i}, \mathcal{E}_{\text{long-term},i}, \underline{d}_{1i}\}_{\forall i}$, longer-term price changes $\mathbf{p_{t-1}} - \mathbf{p_{t-4}}$, exogenous stock characteristics $\mathbf{X_t}$, and unobserved latent demand $\boldsymbol{\epsilon_t}$ extracted from the demand system, to market clearing equilibrium price, based on equation (21).[40] In other words, the function $\mathbf{g}$ determines the equilibrium price p_t that is consistent with individual demand (19), the assets-under-management dynamics (20), and the equilibrium condition (21).

Equation (30) confirms that observed returns are the difference in market clearing prices based on the demand system, which is true by definition of the demand-system estimates:

$$\mathbf{p_t} - \mathbf{p_{t-1}} = \mathbf{g}\left(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}, \mathbf{X_t}, \boldsymbol{\epsilon_t}; \theta\right) - \mathbf{g}\left(\mathbf{p_{t-2}} - \mathbf{p_{t-5}}, \mathbf{X_{t-1}}, \boldsymbol{\epsilon_{t-1}}; \theta\right) \tag{30}$$

$$= \Delta \mathbf{p_t}(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}) + \Delta \mathbf{p_t}(\mathbf{X}) + \Delta \mathbf{p_t}(\epsilon) \tag{31}$$

More importantly, the demand system allows me to trace the contribution of each of the

[40] There are more components to the demand system, for example, the dynamics of asset-under-management. But since they empirically do not contribute to the making of momentum, they have been omitted for brevity.

terms: long-term past returns, stock characteristics, and latent demand, as shown in equation (31). That is, I update each component of the demand system step-by-step and calculate its counterfactual returns:

$$\Delta \mathbf{p_t}(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}) = \mathbf{g}\left(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}, \mathbf{X_{t-1}}, \epsilon_{\mathbf{t-1}}; \theta\right) - \mathbf{g}\left(\mathbf{p_{t-2}} - \mathbf{p_{t-5}}, \mathbf{X_{t-1}}, \epsilon_{\mathbf{t-1}}; \theta\right) \tag{32}$$

$$\Delta \mathbf{p_t}(\mathbf{X}) = \mathbf{g}\left(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}, \mathbf{X_t}, \epsilon_{\mathbf{t-1}}; \theta\right) - \mathbf{g}\left(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}, \mathbf{X_{t-1}}, \epsilon_{\mathbf{t-1}}; \theta\right) \tag{33}$$

$$\Delta \mathbf{p_t}(\epsilon) = \mathbf{g}\left(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}, \mathbf{X_t}, \epsilon_{\mathbf{t}}; \theta\right) - \mathbf{g}\left(\mathbf{p_{t-1}} - \mathbf{p_{t-4}}, \mathbf{X_t}, \epsilon_{\mathbf{t-1}}; \theta\right). \tag{34}$$

Up to this point, I followed Koijen and Yogo (2019) in the definition of counterfactual returns. But next, I form counterfactual momentum portfolio returns to assess which components are responsible for momentum in equilibrium. In particular, I perform standard momentum sorts as of time $t-1$, meaning that I sort stocks into tercile portfolios based on their performance during the formation period, which is 4 to 15 months before t. Then, within each momentum-signal tercile, I calculate value-weighted portfolio returns and calculate the long-short of past winners minus past losers between $t-1$ and t. Based on $t-1$ momentum sorts I calculate long-short returns based on the observed capital gains $\Delta \mathbf{p_t}$ — observed momentum — and also for each of $\Delta \mathbf{p_t}(\mathbf{p_{t-1}} - \mathbf{p_{t-4}})$, $\Delta \mathbf{p_t}(\mathbf{X})$ and $\Delta \mathbf{p_t}(\epsilon)$, corresponding to the portion of momentum driven by the term-structure of demand elasticities, fundamentals, and demand shocks, respectively.

Table 4 implements the decomposition. First, the term structure of demand elasticities is the primary driver of momentum between 1999 and 2020. On its own, it would have generated annualized momentum returns of about 24%. Investors, in aggregate, are more responsive to recent returns than longer-term variation in prices. They respond relatively more elastically to a shock over a horizon of one quarter, limiting its impact on prices. However, as investors subsequently become less willing to continue to absorb the shock, its equilibrium price impact increases, creating momentum. Second, demand shocks are generally mean-reverting,

Table 4. Decomposition of momentum returns

Momentum	Decomposition		
Annualized Return (1999-2020)	Elasticities	Fundamentals	Demand Shocks
2.09%	24.65%	22.11%	−45.71%

Table 4 decomposes total annualized momentum returns between 1999 and 2020 into contributions from the term structure of demand elasticities in column 2, the evolution of demand for fundamentals in column 3, and the persistence of demand shocks in column 4. All reported numbers represent annualized momentum returns.

capturing overreaction rather than underreaction and leading to reversal that undoes most momentum originating from the term structure of elasticities. This is consistent with the overall low momentum returns of about 2% over the sample period, consistent with ideas of anomaly attenuation (Chordia, Subrahmanyam, and Tong, 2014). These two observations could potentially be related if the decline of overall momentum profitability is the result of less persistent demand shocks. For example, Martineau (2021) provides evidence that the post-earnings announcement drift (e.g., Bernard and Thomas, 1989, 1990), a common example of underreaction, has recently disappeared. Third, we can split the impact of baseline demand into the part coming from unobserved demand shocks and the demand for stock characteristics. The component capturing momentum from fundamentals strongly contributes toward momentum. This behavior could be driven by fundamental stock characteristics drifts, such as earnings momentum (Chordia and Shivakumar, 2006). But then rational investors should take such drifts in fundamentals into account when forming their beliefs. An alternative explanation is that the demand for characteristics and latent demand are related, generating a specific form of underreaction. Similar to Novy-Marx (2015), the observed behavior is consistent with past latent demand predicting future stock characteristics that enter investors' demand functions: In the demand system, this mechanism generates both mean-reversion in latent demand and momentum due to investors' demand for fundamentals. Overall, however, mean reversion prevails.

4.2 Demand-system enhanced momentum returns

In the previous section, I showed that the evolution of investor responses to demand shocks is the primary driver of momentum in the cross-section of stock returns. Next, I incorporate this finding into forming a "demand-system enhanced" momentum strategy.

More precisely, I use cross-sectional variation in stock ownership across stocks to predict in which stocks momentum strategies are most profitable.[41] In stocks disproportionally held by momentum-generating investors, that is, investors with downwards-sloping elasticity term structures, the price impact of past shocks gets exacerbated over time. This generates positive serial correlation in stock returns and, thus, stock momentum.

I sort stocks in the cross-section based on their aggregate elasticity term structure $\eta_t(n)$ from section 3.4.1, and then test if momentum strategies' profitability varies based on these sorts. In the context of Table 5, I first sort stocks into two categories based on whether they are above or below the time t cross-sectional median of aggregate stock-level elasticity term-structures, $\eta_t(n)$. Then, within each category, I separately implement momentum strategies. That is, I sort stocks into terciles based on past performance between months $t-12$ to $t-1$, and build portfolios that go long past winners, and short past losers.

Columns 1 to 4 of Table 5 exhibit returns to momentum strategies that value-weight both the long and short legs, while columns 5 to 8 equal-weight returns. The first and fourth column show returns to a standard momentum strategy in all stocks, irrespective of their term structure of elasticities. Momentum returns are generally for the sample period from October 1999 to December 2020: Momentum returns range from an annualized 0 to 5%, depending on whether they are value- or equal weighted, and on whether they average returns or α with respect to standard factor models. These low momentum returns are generally consistent

[41] There are many examples of the importance of ownership structure for returns: Gompers and Metrick (2001) argue that the attenuation of the size premium is partially driven by institutional ownership. Antón and Polk (2014) show that common stock ownership affects stock return correlations. Rzeźnik and Weber (2022) demonstrate that fire sales only generate price pressure in the absence of specialized investors. More generally, intermediary ownership drives returns (Adrian, Etula, and Muir, 2014; He, Kelly, and Manela, 2017; Kargar, 2021), especially for heavily intermediated asset classes (Haddad and Muir, 2021; Eisfeldt et al., 2022).

Table 5. Momentum returns sorted by term structure of elasticities η

All Stocks	Low η	High η	Lo−Hi	All Stocks	Low η	High η	Lo−Hi
Average Returns: Value Weighted				Average Returns: Equally Weighted			
2.09	6.11	−0.72	6.82^{**}	1.87	3.87	0.08	3.79^{**}
(4.04)	(4.26)	(4.31)	(3.43)	(4.41)	(4.62)	(4.23)	(1.77)
Fama-French 3 Factor α: Value Weighted				Fama-French 3 Factor α: Equally Weighted			
5.51^{*}	10.09^{***}	2.45	7.64^{**}	5.33	7.33^{*}	3.77	3.56^{**}
(3.07)	(2.98)	(3.88)	(3.68)	(3.76)	(3.99)	(3.56)	(1.81)
Carhart 4 Factor α: Value Weighted				Carhart 4 Factor α: Equally Weighted			
0.40	4.86	−2.16	7.03^{*}	0.41	2.16	−0.83	2.98
(1.82)	(3.56)	(1.76)	(4.07)	(1.57)	(2.31)	(1.24)	(2.07)

Table 5 reports the returns to momentum strategies, where the long leg consists of the tercile of winners during the formation period and the short leg of the tercile of losers during the formation period. The four left columns report the returns to value-weighted momentum portfolios, while the four right columns use equal weighting. Columns 1 and 5 look at the performance of momentum among all stocks. Columns 2 and 6 filter to stocks with a term structure of elasticity η that is more steeply decreasing than the cross-sectional median. Columns 3 and 7 use stocks not used in columns 2 and 6, and columns 4 and 8 report their difference. The first panel reports average returns, while the second and third panels show the anomaly α with respect to the Fama and French (1993) and Carhart (1997) factor models. The sample period is from 1999 to 2020. Standard errors are estimated using Newey-West with 12 lags. $***$, $**$, and $*$ indicate significance at the 1%, 5%, and 10% level, respectively.

with ideas of anomaly attenuation as in Chordia, Subrahmanyam, and Tong (2014).

However, there is substantial variation in momentum returns based on elasticity-term structures $\eta_t(n)$: Momentum returns are more pronounced in low elasticity-term structure stocks (columns 2 and 6) relative to momentum returns based on the entire universe of stocks by an annualized 4% value-weighted (2% value-weighted). Consistent with this, momentum returns are only economically and statistically significant within low η stocks after controlling for risk as captured by the Fama-French 3 Factor model (Fama and French, 1993). This corresponds to the idea of an "enhanced momentum strategy": Instead of implementing a momentum strategy based on the past performance of the entire universe of stocks, limiting the universe of stocks to those that are more prone to exhibit momentum – low elasticity term-structure stocks – produces superior risk-adjusted performance.

Across specifications, stocks with low η, i.e., stocks with more decreasing elasticity-term structures, have higher momentum returns by about 7% value-weighted (column 4), and 3.5% equally weighted (column 8). Appendix Table IA.2 shows the robustness of these results in a battery of additional tests involving variations on the construction of momentum and elasticity term-structure portfolios and size controls, which are designed to capture issues related to illiquidity.[42] Moreover, these Lo-Hi differences remain constant irrespective of the choice of factor model they are evaluated against. This finding suggests a new strategy: Going long momentum in low η stocks and going short momentum in high η stocks, which are expected to feature reversal rather than momentum. Conceptually, this idea is similar to combining elements of momentum and reversal strategies. In fact, it is equivalent to combining a momentum strategy in high η "momentum stocks" with a reversal strategy in low η "reversal stocks". However, unlike Asness, Moskowitz, and Pedersen (2013), who combine momentum with long-term reversal, I separate stocks based on their momentum- or reversal properties at the same horizon.

The finding that the difference in returns of momentum strategies between low and high elasticity- term structure stocks remains constant across choices of factor models is particularly striking in the context of the Carhart 4 factor model (Carhart, 1997), which contains a momentum factor. It suggests that the variation in momentum strategy returns based on elasticity-term structures is not merely the result of recovering stocks with high β_{Mom}, that is, a high factor-beta with respect to the momentum factor, but instead can point at

[42] First, illiquidity and informational efficiency are particularly relevant for small stocks (Lo and MacKinlay, 1990; Jegadeesh and Titman, 1993; Lakonishok, Shleifer, and Vishny, 1994; Hong, Lim, and Stein, 2000). To see the impact of small stocks, one of the robustness checks in Appendix Table IA.2 looks at the profitability of momentum across the size distribution and shows that my results are robust to conditioning on size. Second, Haddad, Huebner, and Loualiche (2022) show that elasticities are empirically related to measures of liquidity: stocks with low elasticities tend to be more illiquid. This raises the concern that dividing by the aggregate real-time elasticity in equation (27) emphasizes illiquid stocks. Consequently, one of the robustness checks in Appendix Table IA.2 considers sorting on the absolute instead of the relative difference between $\bar{\mathcal{E}}_{\text{recent},t}(n)$ and $\bar{\mathcal{E}}_{\text{long-term},t}(n)$, which does not affect results.

variation in momentum profitability that remains unspanned by the momentum factor itself. Specifically, the factor-beta of the Lo-Hi strategy, or equivalently, the difference in momentum factor exposures between momentum strategies in low versus high η stocks, is close to zero, at $\beta_{Mom} = 0.12$.

Momentum strategies are known to suffer from momentum crashes (Daniel and Moskowitz, 2016), periods during which momentum performs exceptionally poorly. If the high returns of the proposed enhanced momentum strategy were driven by high factor-betas with respect to the momentum factor, then the strategy would necessarily suffer from momentum crashes as well. In fact, its momentum crashes would be proportionally more severe. As it is, that need not be the case. Below, I examine the performance of the proposed enhanced strategies during times when traditional momentum strategies crash.

4.2.1 Momentum crashes

Daniel and Moskowitz (2016) identify two prolonged periods they label momentum crashes, following the Great Depression (June 1932 to December 1939) and the 2008-2009 financial crisis (March 2009 to March 2013). I study the performance of enhanced momentum strategies during the latter of these two momentum crashes.[43]

Table 6 is equivalent to Table 5, but zooms into the momentum crash period from March 2009 to March 2013. Columns 1 and 4 show that average annualized momentum returns based on the full universe of stocks were low, at about -7.5% equal-weighted and -9% value-weighted. Accounting for factor exposures accounts for most of this negative performance.

Implementing a momentum strategy in low elasticity term structure stocks would have largely avoided the momentum crash. Most strikingly, the gap in momentum performance between low and high η stocks during momentum crashes is particularly wide at an annualized

[43] One caveat for the results of this section is that I study the only large momentum crash that occurred during my already relatively short sample period. Consequently, results may not be representative of other momentum crashes. Nevertheless, as Table 6 shows, the difference between momentum performance in low- and high-term-structure of elasticity stocks is strongly statistically significant, despite the short sample period.

Table 6. Momentum returns sorted by term structure of elasticities η during the March 2009 to March 2013 momentum crash

All Stocks	Low η	High η	Lo−Hi	All Stocks	Low η	High η	Lo−Hi
Average Returns: Value Weighted				Average Returns: Equally Weighted			
−8.97	−3.48	−12.35	8.88***	−7.61	−3.85	−11.19	7.33***
(11.94)	(11.45)	(11.55)	(3.06)	(14.84)	(14.28)	(15.10)	(1.97)
Fama-French 3 Factor α: Value Weighted				Fama-French 3 Factor α: Equally Weighted			
−0.45	7.39	−5.17	12.56***	−1.16	3.26	−4.57	7.82***
(6.05)	(5.60)	(6.46)	(4.36)	(8.85)	(8.94)	(8.60)	(1.83)
Carhart 4 Factor α: Value Weighted				Carhart 4 Factor α: Equally Weighted			
2.23	10.01**	−2.83**	12.84***	1.27	5.43	−1.97	7.40***
(2.51)	(4.60)	(1.11)	(4.74)	(3.18)	(4.09)	(2.31)	(2.29)

Table 6 reports the returns to momentum strategies from March 2009 to March 2013, a momentum crash period identified by Daniel and Moskowitz (2016). Besides filtering to a period of momentum crashing, the construction of the table is equivalent to table 5: The left four columns report the returns to value-weighted momentum portfolios, while the right four columns use equal weighting. Columns 1 and 5 look at the performance of momentum among all stocks. Columns 2 and 6 filter to stocks with a term structure of elasticity η that is more steeply decreasing than the cross-sectional median. Columns 3 and 7 use stocks not used in columns 2 and 6, and columns 4 and 8 report their difference. The first panel reports average returns, while the second and third panels show the anomaly α with respect to the Fama and French (1993) and Carhart (1997) factor models. The sample period is from March 2009 to March 2013. Standard errors are estimated using Newey-West with 12 lags. $***$, $**$, and $*$ indicate significance at the 1%, 5%, and 10% level, respectively.

$9 - 12.5\%$ value-weighted and 7.5% equal-weighted across specifications.

Since momentum crashes typically occur immediately following stock market crashes, they likely coincide with high marginal utility states. This would suggest that unconditional outperformance of momentum in low η stocks could be fully consistent with rational explanations as compensation for momentum-crash-related risk if low η stocks would suffer from particularly strong momentum crashes. Instead, the opposite is the case. Low term-structure of elasticity stocks do not only have larger momentum returns unconditionally; they even have larger momentum returns during times when marginal utility is likely to be high. Relatedly, Daniel and Moskowitz (2016) show that this is true for momentum returns more generally, as crashes are partially predictable, such that timing momentum improves performance.[44]

5 Conclusion

Momentum in stock returns is one of the most widely studied anomalies, with many papers proposing explanations for momentum based on some form of underreaction. In this paper, I emphasize the role of a complementary channel: the term structure of demand elasticities, representing investors' differential responses to short- and longer-term price variation. I put forward a framework incorporating both the direct evolution of demand shocks over time and investors' dynamic reactions to price changes across horizons. Finally, I estimate the model for institutional investors in the U.S. stock market between 1999 and 2020.

My estimates suggest that the main driver of momentum returns is the downward-sloping term structure of elasticities. On average, investors are 25% less responsive to longer-term variation in prices than to recent price changes over the previous quarter. Institutions exceptionally responsive to recent price changes drive this overall pattern. In contrast, demand shocks exhibit mean reversion and thus generate reversal.

My results suggest the need to incorporate dynamic investor responses into models of momentum generation. Yet beyond the application in this paper, differential responses to price changes that are more nuanced across horizons could also help us understand a larger class of price-based anomalies in a unified framework. For example, besides momentum, there are short-term and long-term reversals. A rich term structure of elasticities could reproduce such time-series patterns. Initially, it would be upward-sloping within the first month, potentially capturing inattention (Duffie, 2010). Next comes the effect documented in this paper: Investors respond more strongly to recent price changes than medium-term variation, generating a downward-sloping term structure of elasticities between a month and a

[44] Similarly, Burnside et al. (2011) show that peso problems cannot fully explain the performance of currency carry trades because carry remains profitable after hedging out extreme disaster risk.

year. And finally, long-term value investors step in, which could be captured by an increasing term structure at long horizons beyond one year.

References

Adam, Klaus and Stefan Nagel. 2022. "Expectations Data in Asset Pricing." Tech. rep., National Bureau of Economic Research.

Adrian, Tobias, Erkko Etula, and Tyler Muir. 2014. "Financial Intermediaries and the Cross-Section of Asset Returns." *The Journal of Finance* 69 (6):2557–2596.

Angrist, Joshua D and Alan B Krueger. 1992. "The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples." *Journal of the American statistical Association* 87 (418):328–336.

Antón, Miguel and Christopher Polk. 2014. "Connected Stocks." *The Journal of Finance* 69 (3):1099–1127.

Arellano, Manuel and Costas Meghir. 1992. "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets." *The Review of Economic Studies* 59 (3):537–559.

Asness, Clifford S., Tobias J. Moskowitz, and Lasse Heje Pedersen. 2013. "Value and Momentum Everywhere." *The Journal of Finance* 68 (3):929–985.

Backus, Matthew, Christopher Conlon, and Michael Sinkinson. 2019. "Common ownership in America: 1980-2017." Tech. rep., National Bureau of Economic Research.

Backus, Matthew, Christopher T Conlon, and Michael Sinkinson. 2020. "Common Ownership Data: Scraped SEC form 13F filings for 1999-2017."

Balasubramaniam, Vimal, John Y Campbell, Tarun Ramadorai, and Benjamin Ranish. 2021. "Who Owns What? A Factor Model for Direct Stockholding." Mimeo, Harvard University.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer. 2015. "X-CAPM: An extrapolative capital asset pricing model." *Journal of Financial Economics* 115 (1):1–24.

———. 2018. "Extrapolation and Bubbles." *Journal of Financial Economics* 129 (2):203–227.

Barberis, Nicholas and Andrei Shleifer. 2003. "Style investing." *Journal of Financial Economics* 68 (2):161–199.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny. 1998. "A model of investor sentiment." *Journal of Financial Economics* 49 (3):307–343.

Bastianello, Francesca and Paul Fontanier. 2021. "Partial Equilibrium Thinking in General Equilibrium." Tech. rep., Harvard.

Berk, Jonathan B., Richard C. Green, and Vasant Naik. 1999. "Optimal Investment, Growth Options, and Security Returns." *The Journal of Finance* 54 (5):1553–1607.

Berk, Jonathan B. and Richard C. Green. 2004. "Mutual Fund Flows and Performance in Rational Markets." *Journal of Political Economy* 112 (6):1269–1295.

Bernard, Victor L. and Jacob K. Thomas. 1989. "Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium?" *Journal of Accounting Research* 27:1–36.

———. 1990. "Evidence that stock prices do not fully reflect the implications of current earnings for future earnings." *Journal of Accounting and Economics* 13 (4):305–340.

Bordalo, Pedro, Nicola Gennaioli, Rafael LaPorta, and Andrei Shleifer. 2022. "Belief Overreaction and Stock Market Puzzles." Tech. rep., Working paper.

Brandt, Michael W., Pedro Santa-Clara, and Rossen Valkanov. 2009. "Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns." *The Review of Financial Studies* 22 (9):3411–3447.

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma. 2020. "Institutional corporate bond pricing." Tech. rep., Swiss Finance Institute Research Paper.

Burnside, Craig, Martin Eichenbaum, Isaac Kleshchelski, and Sergio Rebelo. 2011. "Do Peso Problems Explain the Returns to the Carry Trade?" *Review of Financial Studies* 24 (3):853–891.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 2011. "Carry Trade and Momentum in Currency Markets." *Annual Review of Financial Economics* 3 (1):511–535.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52 (1):57–82.

Cassella, Stefano and Huseyin Gulen. 2018. "Extrapolation Bias and the Predictability of Stock Returns by Price-Scaled Variables." *The Review of Financial Studies* 31 (11):4345–4397.

Chan, Louis K. C., Narasimhan Jegadeesh, and Josef Lakonishok. 1996. "Momentum Strategies." *The Journal of Finance* 51 (5):1681–1713.

Charles, Constantin, Cary Frydman, and Mete Kilic. 2022. "Insensitive Investors." Working paper.

Chaudhry, Aditya. 2022. "Do Subjective Growth Expectations Matter for Asset Prices?" Working paper.

Chen, Hsiu-Lang and Werner De Bondt. 2004. "Style momentum within the S&P-500 index." *Journal of Empirical Finance* 11 (4):483–507.

Chevalier, Judith and Glenn Ellison. 1997. "Risk Taking by Mutual Funds as a Response to Incentives." *Journal of Political Economy* 105 (6):1167–1200.

Chordia, Tarun and Lakshmanan Shivakumar. 2006. "Earnings and price momentum." *Journal of Financial Economics* 80 (3):627–656.

Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong. 2014. "Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?" *Journal of Accounting and Economics* 58 (1):41–58.

Chui, Andy C W, Avanidhar Subrahmanyam, and Sheridan Titman. 2022. "Momentum, Reversals, and Investor Clientele." *Review of Finance* 26 (2):217–255.

Collin-Dufresne, Pierre, Michael Johannes, and Lars A. Lochstoer. 2016. "Parameter Learning in General Equilibrium: The Asset Pricing Implications." *American Economic Review* 106 (3):664–98.

Coval, Joshua and Erik Stafford. 2007. "Asset fire sales (and purchases) in equity markets." *Journal of Financial Economics* 86 (2):479–512. URL `https://www.sciencedirect.com/science/article/pii/S0304405X07001158`.

Da, Zhi, Xing Huang, and Lawrence J. Jin. 2021. "Extrapolative beliefs in the cross-section: What can we learn from the crowds?" *Journal of Financial Economics* 140 (1):175–196.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam. 1998. "Investor psychology and security market under-and overreactions." *the Journal of Finance* 53 (6):1839–1885.

Daniel, Kent and Tobias J. Moskowitz. 2016. "Momentum crashes." *Journal of Financial Economics* 122 (2):221–247.

Daniel, Kent D, Alexander Klos, and Simon Rottke. 2021. "The dynamics of disagreement." Tech. rep., National Bureau of Economic Research.

Davis, Carter. 2021. "Machine learning, quantitative portfolio choice, and mispricing." Working paper.

Davis, Carter, Mahyar Kargar, and Jiacui Li. 2022. "An Information-Based Explanation for Inelastic Demand." Working paper.

De Bondt, Werner FM and Richard Thaler. 1985. "Does the stock market overreact?" *The Journal of finance* 40 (3):793–805.

De La O, Ricardo and Sean Myers. 2021. "Subjective Cash Flow and Discount Rate Expectations." *The Journal of Finance* 76 (3):1339–1387.

Dong, Xi, Namho Kang, and Joel Peress. 2022. "Fast and Slow Arbitrage: The Predictive Power of Capital Flows for Factor Returns." Tech. rep.

Dou, Winston, Leonid Kogan, and Wei Wu. 2020. "Common fund flows: Flow hedging and factor pricing." *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper* .

Duffie, Darrell. 2010. "Presidential Address: Asset Price Dynamics with Slow-Moving Capital." *The Journal of Finance* 65 (4):1237–1267.

Ehsani, Sina and Juhani T. Linnainmaa. 2022. "Factor Momentum and the Momentum Factor." *The Journal of Finance* 77 (3):1877–1919.

Eisfeldt, Andrea L, Bernard Herskovic, Sriram Rajan, and Emil Siriwardane. 2022. "OTC Intermediaries." *The Review of Financial Studies* .

Fama, Eugene F. 2014. "Two pillars of asset pricing." *American Economic Review* 104 (6):1467–85.

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

Gabaix, Xavier and Ralph SJ Koijen. 2020. "In search of the origins of financial fluctuations: The inelastic markets hypothesis." Mimeo, Harvard University.

Gabaix, Xavier, Ralph SJ Koijen, Federico Mainardi, Sangmin Oh, and Motohiro Yogo. 2022. "Asset Demand of US Households." Mimeo.

Gabaix, Xavier, Arvind Krishnamurthy, and Olivier Vigneron. 2007. "Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market." *The Journal of Finance* 62 (2):557–595.

Gabaix, Xavier and Matteo Maggiori. 2015. " International Liquidity and Exchange Rate Dynamics." *The Quarterly Journal of Economics* 130 (3):1369–1420.

Gârleanu, Nicolae, Lasse Heje Pedersen, and Allen M. Poteshman. 2009. "Demand-Based Option Pricing." *The Review of Financial Studies* 22 (10):4259–4299.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus. 2021. "Five facts about beliefs and portfolios." *American Economic Review* 111 (5):1481–1522.

Gompers, Paul A. and Andrew Metrick. 2001. "Institutional Investors and Equity Prices." *The Quarterly Journal of Economics* 116 (1):229–259.

Greenwood, Robin, Samuel G Hanson, and Gordon Y Liao. 2018. "Asset Price Dynamics in Partially Segmented Markets." *The Review of Financial Studies* 31 (9):3307–3343.

Greenwood, Robin and Andrei Shleifer. 2014. "Expectations of Returns and Expected Returns." *The Review of Financial Studies* 27 (3):714–746.

Greenwood, Robin and Dimitri Vayanos. 2014. "Bond Supply and Excess Bond Returns." *The Review of Financial Studies* 27 (3):663–713.

Greenwood, Robin M and Annette Vissing-Jorgensen. 2018. "The impact of pensions and insurance on global yield curves." Tech. rep., Harvard Business School Finance Working Paper.

Grinblatt, Mark and Bing Han. 2005. "Prospect theory, mental accounting, and momentum." *Journal of Financial Economics* 78 (2):311–339.

Grinblatt, Mark and Matti Keloharju. 2000. "The investment behavior and performance of various investor types: a study of Finland's unique data set." *Journal of Financial Economics* 55 (1):43–67.

Grinblatt, Mark and Tobias J. Moskowitz. 2004. "Predicting stock price movements from past returns: the role of consistency and tax-loss selling." *Journal of Financial Economics* 71 (3):541–579.

Grinblatt, Mark, Sheridan Titman, and Russ Wermers. 1995. "Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior." *The American Economic Review* 85 (5):1088–1105.

Gromb, Denis and Dimitri Vayanos. 2002. "Equilibrium and welfare in markets with financially constrained arbitrageurs." *Journal of Financial Economics* 66 (2):361–407.

Grossman, Sanford J. and Merton H. Miller. 1988. "Liquidity and Market Structure." *The Journal of Finance* 43 (3):617–633.

Grossman, Sanford J and Joseph E Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70 (3):393–408.

Guibaud, Stéphane, Yves Nosbusch, and Dimitri Vayanos. 2013. "Bond market clienteles, the yield curve, and the optimal maturity structure of government debt." *The Review of Financial Studies* 26 (8):1914–1961.

Haddad, Valentin, Paul Huebner, and Erik Loualiche. 2022. "How Competitive is the Stock Market? Theory, Evidence from Portfolios, and Implications for the Rise of Passive Investing." Working paper.

Haddad, Valentin and Tyler Muir. 2021. "Do Intermediaries Matter for Aggregate Asset Prices?" *Journal of Finance* 76 (6):2719–2761.

He, Zhiguo, Bryan Kelly, and Asaf Manela. 2017. "Intermediary asset pricing: New evidence from many asset classes." *Journal of Financial Economics* 126 (1):1–35.

Hong, Harrison, Terence Lim, and Jeremy C. Stein. 2000. "Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies." *The Journal of Finance* 55 (1):265–295.

Hong, Harrison and Jeremy C Stein. 1999. "A unified theory of underreaction, momentum trading, and overreaction in asset markets." *The Journal of Finance* 54 (6):2143–2184.

Ippolito, Richard A. 1992. "Consumer reaction to measures of poor quality: Evidence from the mutual fund industry." *The Journal of Law and Economics* 35 (1):45–70.

Jansen, Kristy AE. 2021. "Long-term Investors, Demand Shifts, and Yields." Tech. rep.

Jegadeesh, Narasimhan and Sheridan Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *The Journal of Finance* 48 (1):65–91.

———. 2001. “Profitability of Momentum Strategies: An Evaluation of Alternative Explanations.” *The Journal of Finance* 56 (2):699–720.

Jiang, Zhengyang, Robert Richmond, and Tony Zhang. 2020. “A portfolio approach to global imbalances.” Mimeo.

Jiang, Zhengyang, Robert J Richmond, and Tony Zhang. 2022. “Understanding the Strength of the Dollar.” Mimeo.

Johnson, Timothy C. 2002. “Rational Momentum Effects.” *The Journal of Finance* 57 (2):585–608.

Kargar, Mahyar. 2021. “Heterogeneous intermediary asset pricing.” *Journal of Financial Economics* 141 (2):505–532.

Koijen, Ralph S. J. and Motohiro Yogo. 2019. “A Demand System Approach to Asset Pricing.” *Journal of Political Economy* 127 (4):1475–1515.

Koijen, Ralph S.J., François Koulischer, Benoît Nguyen, and Motohiro Yogo. 2021. “Inspecting the mechanism of quantitative easing in the euro area.” *Journal of Financial Economics* 140 (1):1–20.

Koijen, Ralph SJ, Robert J Richmond, and Motohiro Yogo. 2020. “Which Investors Matter for Equity Valuations and Expected Returns?” Mimeo, National Bureau of Economic Research.

Koijen, Ralph SJ and Motohiro Yogo. 2020. “Exchange rates and asset prices in a global demand system.” Mimeo, National Bureau of Economic Research.

Kyle, Albert S. 1985. “Continuous Auctions and Insider Trading.” *Econometrica* 53 (6):1315–1335. URL http://www.jstor.org/stable/1913210.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny. 1994. “Contrarian Investment, Extrapolation, and Risk.” *The Journal of Finance* 49 (5):1541–1578.

Lo, Andrew W. and A. Craig MacKinlay. 1990. “When Are Contrarian Profits Due to Stock Market Overreaction?” *The Review of Financial Studies* 3 (2):175–205.

Lochstoer, Lars A and Tyler Muir. 2022. “Volatility expectations and returns.” *The Journal of Finance* 77 (2):1055–1096.

Long, J. Bradford De, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. 1990. “Positive Feedback Investment Strategies and Destabilizing Rational Speculation.” *The Journal of Finance* 45 (2):379–395.

Lou, Dong. 2012. “A Flow-Based Explanation for Return Predictability.” *The Review of Financial Studies* 25 (12):3457–3489.

Lou, Dong and Christopher Polk. 2021. “Comomentum: Inferring Arbitrage Activity from Return Correlations.” *The Review of Financial Studies* 35 (7):3272–3302.

Luo, Jiang, Avanidhar Subrahmanyam, and Sheridan Titman. 2020. "Momentum and Reversals When Overconfident Investors Underestimate Their Competition." *The Review of Financial Studies* 34 (1):351–393.

Malmendier, Ulrike and Stefan Nagel. 2011. "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?*." *The Quarterly Journal of Economics* 126 (1):373–416.

———. 2016. " Learning from Inflation Experiences *." *The Quarterly Journal of Economics* 131 (1):53–87.

Martineau, Charles. 2021. "Rest in peace post-earnings announcement drift." *Critical Finance Review* .

McCarthy, Odhrain and Sebastian Hillenbrand. 2021. "Heterogeneous Investors and Stock Market Fluctuations." *Available at SSRN 3944887* .

Menkhoff, Lukas, Lucio Sarno, Maik Schmeling, and Andreas Schrimpf. 2012. "Currency momentum strategies." *Journal of Financial Economics* 106 (3):660–684.

Merton, Robert C. 1987. "A Simple Model of Capital Market Equilibrium with Incomplete Information." *The Journal of Finance* 42 (3):483–510.

Mitchell, Mark, Lasse Heje Pedersen, and Todd Pulvino. 2007. "Slow moving capital." *American Economic Review* 97 (2):215–220.

Moskowitz, Tobias J. and Mark Grinblatt. 1999. "Do Industries Explain Momentum?" *The Journal of Finance* 54 (4):1249–1290. URL `http://www.jstor.org/stable/798005`.

Moskowitz, Tobias J., Yao Hua Ooi, and Lasse Heje Pedersen. 2012. "Time series momentum." *Journal of Financial Economics* 104 (2):228–250.

Nagel, Stefan and Zhengyang Xu. 2022. "Asset pricing with fading memory." *The Review of Financial Studies* 35 (5):2190–2245.

Noh, Don and Sangmin Oh. 2020. "Measuring institutional pressure for greenness: A demand system approach." Mimeo.

Novy-Marx, Robert. 2012. "Is momentum really momentum?" *Journal of Financial Economics* 103 (3):429–453.

———. 2015. "Fundamentally, momentum is fundamental momentum." Tech. rep., National Bureau of Economic Research.

Pastor, Lubos and Robert F. Stambaugh. 2003. "Liquidity Risk and Expected Stock Returns." *Journal of Political Economy* 111 (3):642–685.

Pavlova, Anna and Taisiya Sikorskaya. 2022. "Benchmarking Intensity." *The Review of Financial Studies* .

Petajisto, Antti. 2009. “Why Do Demand Curves for Stocks Slope Down?” *Journal of Financial and Quantitative Analysis* 44 (5):1013–1044.

Rzeźnik, Aleksandra and Rüdiger Weber. 2022. “Money in the Right Hands: The Price Effects of Specialized Demand.” Tech. rep.

Sadka, Ronnie. 2006. “Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk.” *Journal of Financial Economics* 80 (2):309–349.

Shleifer, Andrei and Robert W. Vishny. 1997. “The Limits of Arbitrage.” *The Journal of Finance* 52 (1):35–55.

Siriwardane, Emil, Aditya Sunderam, and Jonathan Wallen. 2021. “Segmented Arbitrage.” Tech. rep., Working paper, HBS.

Sirri, Erik R. and Peter Tufano. 1998. “Costly Search and Mutual Fund Flows.” *The Journal of Finance* 53 (5):1589–1622.

van der Beck, Philippe. 2021. “Flow-driven ESG returns.” Tech. rep., Swiss Finance Institute Research Paper.

———. 2022. “On the Estimation of Demand-Based Asset Pricing Models.” Tech. rep., Swiss Finance Institute.

Vayanos, Dimitri and Jean-Luc Vila. 2021. “A Preferred-Habitat Model of the Term Structure of Interest Rates.” *Econometrica* 89 (1):77–112.

Veldkamp, Laura L. 2011. *Information Choice in Macroeconomics and Finance.* Princeton University Press.

A Equilibrium Momentum from Dynamic Trading

This appendix details formal derivations for Section 2. In particular, the focus of this section is on deriving the results of Proposition 1 in the presence of heterogeneous investors as detailed in Section 2.4. That is, investors have demand curves

$$d_{it} = \underline{d}_i - \mathcal{E}_{\text{recent},i} \times (p_t - p_{t-1}) - \mathcal{E}_{\text{long-term},i} \times (p_{t-1} - p_{t-s}) \tag{IA.1}$$

$$D_t^N = \phi \times D_{t-1}^N + \epsilon_t^N, \tag{IA.2}$$

where lower-case and upper-case letters represent logs and levels, respectively. Here i denotes an investor with elasticity $\mathcal{E}_{\text{recent},i}$ to recent and $\mathcal{E}_{\text{long-term},i}$ to longer-term price changes. Investor N has persistent demand with demand shock ϵ_t^N and persistence ϕ.

Next, define the aggregate, holdings-weighted recent and longer-term elasticities $\bar{\mathcal{E}}_{\text{recent},t}$ and $\bar{\mathcal{E}}_{\text{long-term},t}$:

$$\bar{\mathcal{E}}_{\text{recent},t} \equiv \int \exp(d_{it})\mathcal{E}_{\text{recent},i} di \tag{IA.3}$$

$$\bar{\mathcal{E}}_{\text{long-term},t} \equiv \int \exp(d_{it})\mathcal{E}_{\text{long-term},i} di. \tag{IA.4}$$

Based on fixed supply S, the market-clearing equation is:

$$\int D_{it} di = \int \exp(d_{it}) di = S - D_t^N. \tag{IA.5}$$

The model of Section 2.1 represents a special case of this setup with $I = 2$ investors: investor ST with $\underline{d}_{ST} = \underline{d}^{ST}$, $\mathcal{E}_{\text{recent},ST} = \mathcal{E}_{\text{recent}}$, $\mathcal{E}_{\text{long-term},ST} = 0$, and investor LT with $\underline{d}_{LT} = \underline{d}^{LT}$, $\mathcal{E}_{\text{recent},LT} = 0$, $\mathcal{E}_{\text{long-term},LT} = \mathcal{E}_{\text{long-term}}$. Section 2.2 further collapses these two investors into one, and Section 2.3 sets $\phi = 1$. Below, I only provide derivations for the general case with heterogenous investors and $\phi \geq 0$ as in Section 2.4. All results in previous sections follow directly.

A.1 Derivations underlying the price impact of a recent demand shock

How much do prices move when a demand shock ϵ_t^N arrives in the market? The answer depends on how strongly investors respond to recent price changes and, specifically, is proportional to the inverse of the aggregate elasticity, $\bar{\mathcal{E}}_{\text{recent},t}^{-1}$. Below I show the derivations behind this result.

Start with an exogenous shock to demand, ϵ_t^N. Such a shock moves the effective supply of the asset, and consequently, the price of the asset increases. Differentiating both sides of the market-clearing equation (IA.5):

$$\frac{d}{d\epsilon_t^N}\int \exp(d_{it})di = -\int \exp(d_{it})\mathcal{E}_{\text{recent},i}\frac{d\Delta p_t}{d\epsilon_t^N}di = -\bar{\mathcal{E}}_{\text{recent},t}\frac{d\Delta p_t}{d\epsilon_t^N} = -1 = \frac{d}{d\epsilon_t^N}\left(S - D_t^N\right). \tag{IA.6}$$

The immediate price impact of a demand shock is:

$$\frac{d\Delta p_t}{d\epsilon_t^N} = \bar{\mathcal{E}}_{\text{recent},t}^{-1}. \tag{IA.7}$$

Define effective supply $\tilde{S}_t$ as $\tilde{S}_t \equiv S - D_t^N$. Then equation (11) follows. That is, one unit of an effective supply shock moves prices by the inverse aggregate recent elasticity, $\bar{\mathcal{E}}_{\text{recent},t}^{-1}$.

A.2 Derivations underlying the dynamic price impact

Now move forward one period. Is there a follow-on price impact to a shock to a demand shock from the previous period? Again start with an exogenous demand shock, ϵ_{t-1}^N, but already occurring at time $t-1$, such that it moves prices at $t-1$. Based on the market-clearing equation (IA.5):

$$\frac{d}{d\epsilon_{t-1}^N}\int \exp(d_{it})di = -\frac{dD_t^N}{d\epsilon_{t-1}^N} \tag{IA.8}$$

$$-\int \exp(d_{it})\left(\mathcal{E}_{\text{recent},i}\frac{d\Delta p_t}{d\epsilon_{t-1}^N} + \mathcal{E}_{\text{long-term},i}\frac{d\Delta p_{t-1}}{d\epsilon_{t-1}^N}\right)di = \phi\frac{dD_{t-1}^N}{d\epsilon_{t-1}^N} \tag{IA.9}$$

$$-\bar{\mathcal{E}}_{\text{recent},t}\frac{d\Delta p_t}{d\epsilon_{t-1}^N} - \bar{\mathcal{E}}_{\text{long-term},t}\frac{d\Delta p_{t-1}}{d\epsilon_{t-1}^N} = \phi \tag{IA.10}$$

$$\frac{\bar{\mathcal{E}}_{\text{recent},t}\frac{d\Delta p_t}{d\epsilon_{t-1}^N} + \bar{\mathcal{E}}_{\text{long-term},t}\frac{d\Delta p_{t-1}}{d\epsilon_{t-1}^N}}{\bar{\mathcal{E}}_{\text{recent},t-1}\frac{d\Delta p_{t-1}}{d\epsilon_{t-1}^N}} = \phi \tag{IA.11}$$

$$\frac{\bar{\mathcal{E}}_{\text{recent},t}\frac{d\Delta p_t}{d\epsilon_{t-1}^N} + \bar{\mathcal{E}}_{\text{long-term},t}\frac{d\Delta p_{t-1}}{d\epsilon_{t-1}^N}}{\bar{\mathcal{E}}_{\text{recent},t}\frac{d\Delta p_{t-1}}{d\epsilon_{t-1}^N}} = \phi\frac{\bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}} \tag{IA.12}$$

$$\frac{\frac{d\Delta p_t}{d\epsilon_{t-1}^N}}{\frac{d\Delta p_{t-1}}{d\epsilon_{t-1}^N}} + \frac{\bar{\mathcal{E}}_{\text{long-term},t} - \bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}} = (\phi - 1)\frac{\bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}}. \tag{IA.13}$$

Rearranging leads to the follow-on price impact of a past demand shocks, as displayed in equations (16) and (17):

$$\frac{d\Delta p_t}{d\epsilon^N_{t-1}} = \left((\phi - 1) \frac{\bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}} - \frac{\bar{\mathcal{E}}_{\text{long-term},t} - \bar{\mathcal{E}}_{\text{recent},t-1}}{\bar{\mathcal{E}}_{\text{recent},t}} \right) \frac{d\Delta p_{t-1}}{d\epsilon^N_{t-1}} \tag{IA.14}$$

$$\approx \left(\phi - 1 - \frac{\bar{\mathcal{E}}_{\text{long-term},t} - \bar{\mathcal{E}}_{\text{recent},t}}{\bar{\mathcal{E}}_{\text{recent},t}} \right) \frac{d\Delta p_{t-1}}{d\epsilon^N_{t-1}}. \tag{IA.15}$$

The approximation in (IA.15) replaces $\bar{\mathcal{E}}_{\text{recent},t-1}$ with $\bar{\mathcal{E}}_{\text{recent},t}$. It shuts down a second-order effect based on local time-series variation in aggregate recent elasticities from period to period. In my estimates, such variation is solely driven by composition effects in stock ownership. However, a stock's ownership distribution is strongly persistent over time, motivating this approximation.

B Identification Strategy

B.1 Moment Conditions

I estimate the model for each investor i using an instrumental variables approach. The identifying assumption is:

$$\mathbf{E}_i \left[\epsilon_{ikt} | \mathbf{X}_{kt}, \widehat{\Delta p}_{ikt}, \widehat{\Delta p}_{ik,t-1}\right] = 0. \tag{IA.16}$$

The resulting moment conditions are:

$$\begin{aligned}
\mathbf{E}_i \left[\epsilon_{ikt}\right] &= 0, \forall i, \forall t && \text{(IA.17)}\\
\mathbf{E}_i \left[\epsilon_{ikt}\mathbf{X}_{kt}\right] &= \mathbf{0}, \forall i && \text{(IA.18)}\\
\mathbf{E}_i \left[\epsilon_{ikt}\widehat{\Delta p}_{ikt}\right] &= 0, \forall i && \text{(IA.19)}\\
\mathbf{E}_i \left[\epsilon_{ikt}\widehat{\Delta p}_{ik,t-1}\right] &= 0, \forall i && \text{(IA.20)}
\end{aligned}$$

There are precisely as many moment conditions as parameters in the model.

B.2 Deriving flow-induced trading

I start by deriving the flow-induced trading instrument proposed by Lou (2012) by shutting off variation in equilibrium returns from equation (21) that is driven by endogenous sources.

$$\begin{aligned}
\Delta p_{kt} &= \log\left(\frac{\sum_j A_{jt}(\mathbf{\Delta p_t}) w_{jkt}(\Delta p_{kt})}{\sum_j A_{j,t-1} w_{jk,t-1}}\right) && \text{(IA.21)}\\
&\approx \log\left(\frac{\sum_j A_{jt}(\mathbf{\Delta p_t}) w_{jk,t-1}}{\sum_j A_{j,t-1} w_{jk,t-1}}\right) && \text{(IA.22)}\\
&\approx \log\left(\frac{\sum_j A_{j,t-1}(1+f_{jt}) w_{jk,t-1}}{\sum_j A_{j,t-1} w_{jk,t-1}}\right) && \text{(IA.23)}\\
&= \log\left(1 + \frac{\sum_j A_{j,t-1} w_{jk,t-1} f_{jt}}{\sum_j A_{j,t-1} w_{jk,t-1}}\right) && \text{(IA.24)}\\
&\approx \frac{\sum_j A_{j,t-1} w_{jk,t-1} f_{jt}}{\sum_{j\neq i} A_{j,t-1} w_{jk,t-1}} && \text{(IA.25)}\\
&= \sum_j o_{jk,t-1} f_{jt} \equiv FIT_{kt}. && \text{(IA.26)}
\end{aligned}$$

Equation (IA.21) starts with the same market-clearing equation for equilibrium returns as equation (21). Even if we were to exclude investor i, there would be an identification problem from indirect effects by investors moving along their demand curves are excluded by replacing portfolio shares w_{jkt} by past portfolio shares $w_{jk,t-1}$ in equation (IA.22). Indirect effects also operate through wealth effects, which are excluded by replacing institutions' AUM A_{jt} by past their past AUM $A_{j,t-1}$ in equation (IA.23). Equation (IA.25) applies the well-known approximation $\log(1+x) \approx x$, for x close to zero. Finally, equation (IA.26) introduces the instrument, flow-induced trading (FIT_{kt}): The past-ownership weighted average of fund flows. This is a commonly used instrument for returns in the literature (Lou, 2012).

B.3 The identification of elasticities

Are $\mathcal{E}_0^i$ and $\mathcal{E}_1^i$ the elasticities of investor i's demand with respect to current and past returns, respectively?

Denote by q_{ikt} the log number of shares investor i demands of asset k at time t:

$$
\begin{aligned}
q_{ikt} &= \log\left(A_{it} w_{ikt}\right) - p_{kt} && \text{(IA.27)}\\
&= \log\left(A_{it} w_{ikt}\right) - \sum_{s \geq 0} \Delta p_{k,t-s}. && \text{(IA.28)}
\end{aligned}
$$

The contemporaneous return elasticity of demand is:

$$
\begin{aligned}
-\frac{dq_{ikt}}{d\Delta p_{kt}} &= 1 - \frac{d}{d\Delta p_{kt}} \log\left(A_{it} w_{ikt}\right) && \text{(IA.29)}\\
&= 1 - \frac{1}{A_{it} w_{ikt}} \left(A_{it} \underbrace{\frac{dw_{ikt}}{d\Delta p_{kt}}}_{=w_{ikt}(1-\mathcal{E}_0^i)(1-w_{ikt})} + w_{ikt} \underbrace{\frac{dA_{it}}{d\Delta p_{kt}}}_{=A_{i,t-1} w_{ik,t-1}} \right) && \text{(IA.30)}\\
&= 1 - (1-\mathcal{E}_0^i)(1 - \underbrace{w_{ikt}}_{\approx 0}) - \frac{A_{i,t-1}}{A_{it}} \underbrace{w_{ik,t-1}}_{\approx 0} && \text{(IA.31)}\\
&\approx \mathcal{E}_0^i. && \text{(IA.32)}
\end{aligned}
$$

Similar to Koijen and Yogo (2019), the elasticity implied by the demand equation of investor i is not exactly $\mathcal{E}_0^i$, but approximately equal to $\mathcal{E}_0^i$ for present and past portfolio weights close to zero, which is the empirically relevant case. While the factor $1 - w_{ikt}$ on $1 - \mathcal{E}_0^i$ comes from substitution through the outside asset as in their demand system, the term $\frac{A_{i,t-1}}{A_{it}} w_{ik,t-1}$ is unique to this setup: it captures a wealth effect from the dynamics of an institution's assets.

The past return elasticity of demand is:

$$-\frac{dq_{ikt}}{d\Delta p_{k,t-1}} = 1 - \frac{d}{d\Delta p_{k,t-1}} \log\left(A_{it} w_{ikt}\right) \tag{IA.33}$$

$$= 1 - (1 - \mathcal{E}_1^i)(1 - \underbrace{w_{ikt}}_{\approx 0}) - \frac{A_{i,t-2}}{A_{i,t-1}} \underbrace{w_{ik,t-2}}_{\approx 0} \tag{IA.34}$$

$$\approx \mathcal{E}_1^i. \tag{IA.35}$$

A similar derivation as for contemporaneous returns shows that the past return elasticity implied by the model is approximately $\mathcal{E}_1^i$.

Investors are less elastic in stocks they hold a lot of because high returns carry a bigger wealth effect.

C Appendix Tables

Table IA.1. Fund flow persistence and flow-performance relationship

	Quarterly Fund Flow f_{it}				
	(1)	(2)	(3)	(4)	(5)
Lagged Fund Flow $f_{i,t-1}$	0.222***	0.219***	0.218***	0.120**	0.127***
	(0.037)	(0.037)	(0.037)	(0.039)	(0.037)
Lagged Fund Flow $f_{i,t-2}$	0.135***	0.137***	0.136***	0.065**	0.072**
	(0.026)	(0.026)	(0.026)	(0.024)	(0.023)
Lagged Fund Flow $f_{i,t-3}$	0.089***	0.091***	0.090***	0.045***	0.052***
	(0.021)	(0.021)	(0.021)	(0.013)	(0.013)
Lagged Fund Flow $f_{i,t-4}$	0.059**	0.059**	0.059**	0.026*	0.032**
	(0.019)	(0.020)	(0.020)	(0.012)	(0.012)
Lagged Fund Return $\Delta p_{i,t-1}$	0.035*	0.150***	0.157***	0.164***	0.176***
	(0.016)	(0.027)	(0.027)	(0.035)	(0.030)
Lagged Fund Return $\Delta p_{i,t-2}$	0.013	0.045	0.053*	0.078***	0.092***
	(0.014)	(0.026)	(0.027)	(0.021)	(0.021)
Lagged Fund Return $\Delta p_{i,t-3}$	-0.002	0.011	0.020	0.049*	0.064***
	(0.013)	(0.020)	(0.020)	(0.020)	(0.018)
Lagged Fund Return $\Delta p_{i,t-4}$	0.008	-0.013	-0.004	0.035*	0.047**
	(0.015)	(0.021)	(0.021)	(0.016)	(0.016)
Date Fixed Effects		Yes	Yes	Yes	Yes
Size Decile Fixed Effects			Yes		Yes
Fund Fixed Effects				Yes	Yes
N	203,222	203,222	203,222	203,222	203,222
R^2	0.158	0.173	0.179	0.257	0.281

Table IA.1 reports coefficients from a panel regression of quarterly fund flows f_{it} on past fund flows $f_{i,t-s}$ and past fund returns $\Delta p_{i,t-s}$, for s between 1 and 4 quarters. Column 2 adds date-fixed effects. Column 3 adds size-decile fixed effects: Funds are sorted into deciles based on funds' past quarter's fund size, i.e. its total net assets. Column 4 uses date-fixed effects and fund-fixed effects. Column 5 combines all three types of fixed effects. The sample period is 1999-2020. Standard errors are 2-way clustered by date and fund for all columns.

Table IA.2.
Robustness of momentum returns based on the term-structure of elasticities

	Lo-Hi η of Value-Weighted Momentum Returns		
	Average	Fama-French 3 α	Carhart 4 α
(1) Baseline Specification	6.82^{**}	7.64^{**}	7.03^{*}
(2) Momentum Deciles	12.73^{**}	13.23^{**}	12.64^{**}
(3) Elasticity Term-Structure η Quintiles	6.14	8.37^{*}	6.14^{*}
(4) Portfolio Sorts with Size Controls	3.73^{*}	4.05^{*}	3.67
(5) Absolute Elasticity Differences	6.79^{**}	7.41^{**}	6.67^{*}
(6) BE-based Instrument	6.92^{**}	6.57^{**}	6.43^{*}

Table IA.2 reports the difference of value-weighted momentum returns in stocks with a steeply decreasing term structure of elasticities, i.e. stocks with η lower than the median, versus in stocks with a flatter term structure. Column 1 reports average returns, while columns 2 and 3 show the anomaly α with respect to the Fama and French (1993) and Carhart (1997) factor models. Specification (1) is the baseline specification from column 4 of table 5. The baseline specification uses the top tercile of winners during the formation period for the long leg, and the bottom tercile for the short leg. Specification (2) instead defines the long and short legs at the top and bottom deciles. While the baseline specification sorts stock based on whether the term structure of elasticities η is above or below the median, specification (3) contrasts the performance of momentum across η quintiles. Specification (4) non-linearly controls for size by initially sorting stocks by size quintiles and subsequently averaging across them. Specification (5) considers the absolute instead of the relative difference between aggregate real-time and past elasticities, i.e. instead of sorting by η as defined in equation (29), it initially sorts by the difference of the elasticities in equations (27) and (28). Finally, specification (6) uses an alternative instrument that uses book-equity-based pseudo holdings in the construction of the instrument. The sample period is from 1999 to 2020. Standard errors are estimated using Newey-West with 12 lags. $***$, $**$, and $*$ indicate significance at the 1%, 5%, and 10% level, respectively.

D Appendix Figures

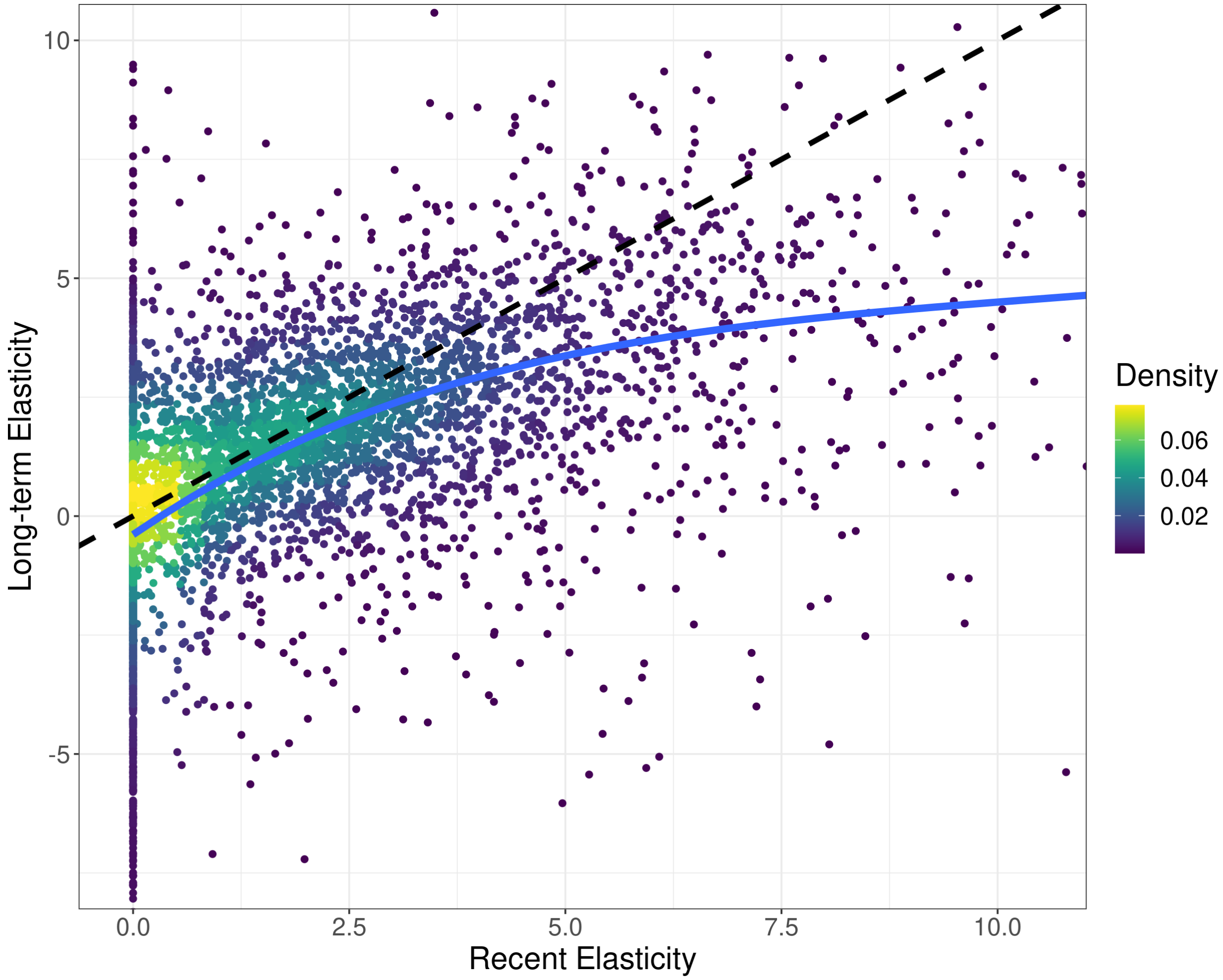

Figure IA.1. Estimates for elasticities $\mathcal{E}_{\text{recent},i}$ and $\mathcal{E}_{\text{long-term},i}$ among institutions with long data histories

Figure IA.1 shows a scatterplot of elasticity estimates for elasticities to price changes over the past quarter, $\mathcal{E}_{\text{recent},i}$, on the x-axis, and variation over the three preceding quarters, $\mathcal{E}_{\text{long-term},i}$, on the y-axis. Compared to Figure 1, it filters to institutions that appear in the data for at least 30 quarters throughout the sample period between 1999 and 2020. Each dot represents one institutional investor in the sample. The solid blue line is a fitted trend line based on cubic regression, and the black dashed line represents flat term structures of elasticities, $\mathcal{E}_{\text{long-term},i} = \mathcal{E}_{\text{recent},i}$. Dots below the dashed line represent downward-sloping term structures of elasticities. The estimation equation is equation (19).

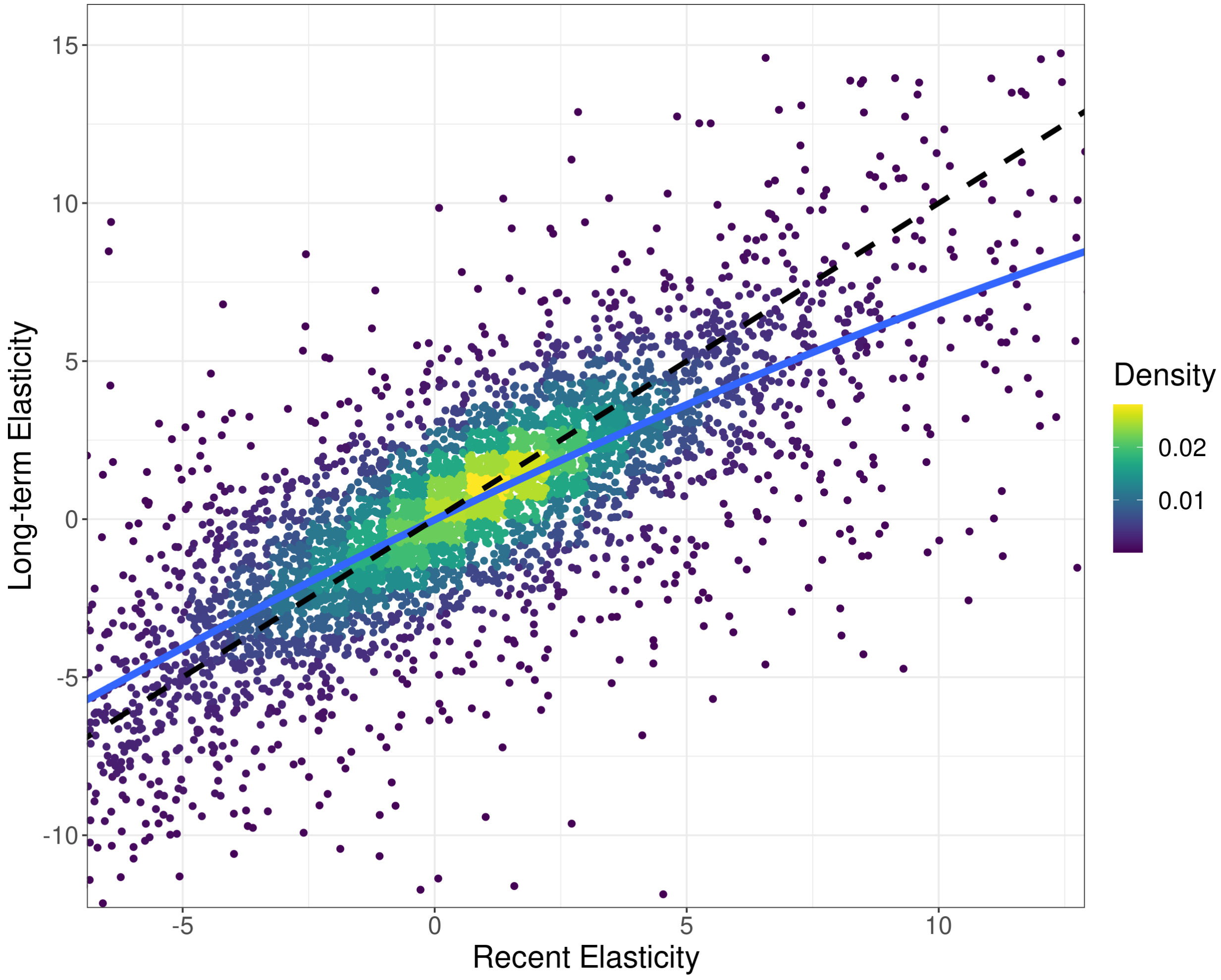

Figure IA.2. Unconstrained estimates for elasticities $\mathcal{E}_{\textbf{recent},i}$ and $\mathcal{E}_{\textbf{long-term},i}$ controlling for the past price level

Figure IA.2 shows a scatterplot of elasticity estimates for elasticities to price changes over the past quarter, $\mathcal{E}_{\text{recent},i}$, on the x-axis, and variation over the three preceding quarters, $\mathcal{E}_{\text{long-term},i}$, on the y-axis. Compared to Figure 1, it allows for (i) negative elasticities to price changes over the previous quarter and (ii) controls for the market-to-book ratio one year ago, instrumented by a Koijen and Yogo (2019) type of instrument. Each dot represents one institutional investor in the sample. The solid blue line is a fitted trend line based on cubic regression, and the black dashed line represents flat term structures of elasticities, $\mathcal{E}_{\text{long-term},i} = \mathcal{E}_{\text{recent},i}$. Dots below the dashed line represent downward-sloping term structures of elasticities. The estimation equation is:

$$\log \frac{w_{it}(n)}{w_{it}(0)} = (1 - \mathcal{E}_{\text{recent},i})\ \Delta p_t(n) + (1 - \mathcal{E}_{\text{long-term},i})\ \left(\sum_{s=1}^{3} \Delta p_{t-s}(n) \right) + (1 - \mathcal{E}_{\text{KY},i}))\ (p_{t-s}(n) - \text{be}_{t-s}(n)) + \underline{d}_{0it} + \underline{d}'_{1i} X_t(n) + \epsilon_{it}(n). \quad \text{(IA.36)}$$

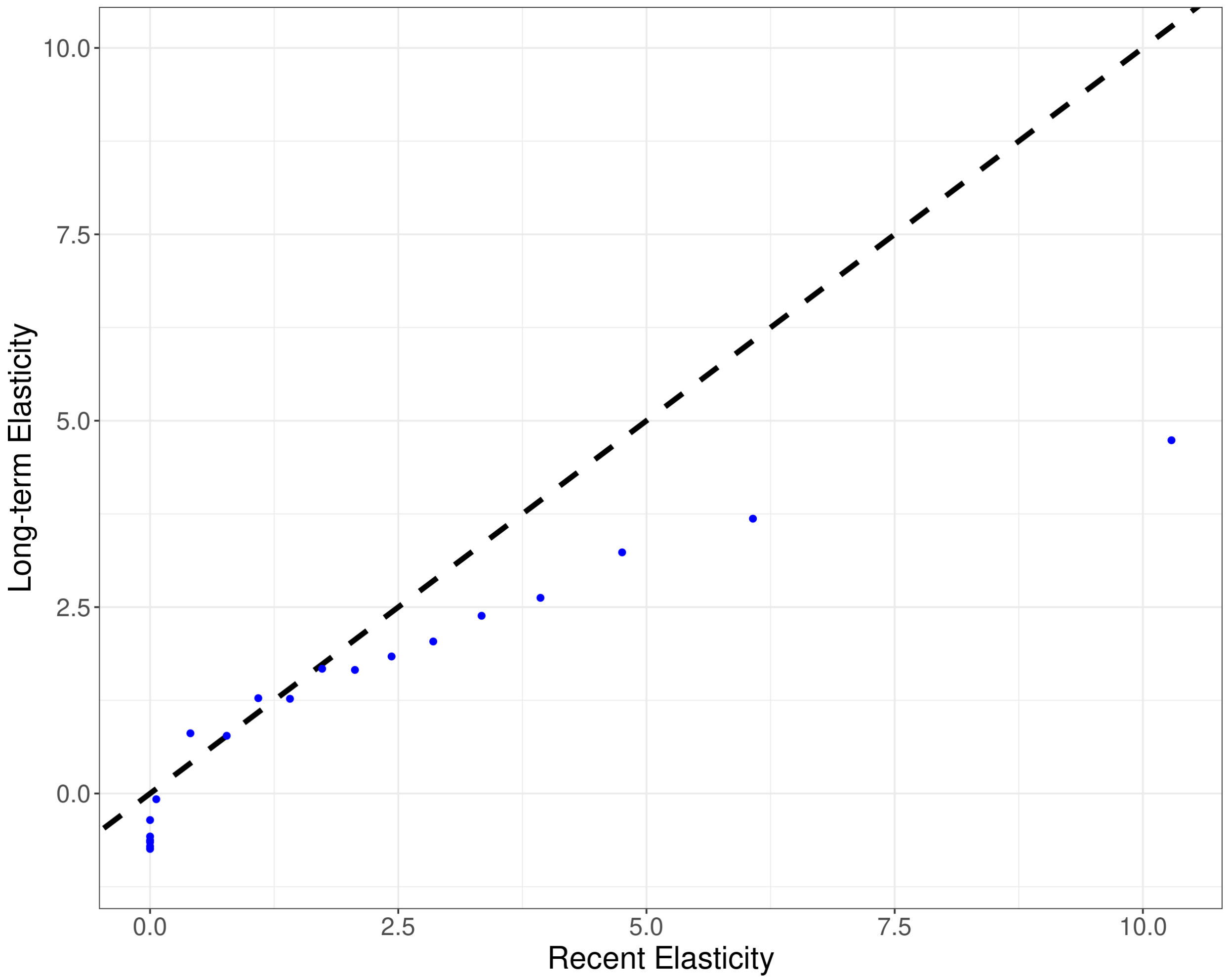

Figure IA.3. Grouped estimates for elasticities $\mathcal{E}_{\textbf{recent},i}$ and $\mathcal{E}_{\textbf{long-term},i}$
Figure IA.3 shows a binned scatterplot of elasticity estimates for elasticities to price changes over the past quarter, $\mathcal{E}_{\text{recent},i}$, on the x-axis, and variation over the three preceding quarters, $\mathcal{E}_{\text{long-term},i}$, on the y-axis. That is, it shows a binned version of Figure 1. Each dot represents one of twenty bins of institutional investors in the sample. The black dashed line represents flat term structures of elasticities, $\mathcal{E}_{\text{long-term},i} = \mathcal{E}_{\text{recent},i}$. Dots below the dashed line represent downward-sloping term structures of elasticities. The estimation equation is equation (19).