The Review of Financial Studies

Asset Price Dynamics with Limited Attention

Terrence Hendershott
University of California at Berkeley

Albert J. Menkveld
Vrije Universiteit Amsterdam

Rémy Praz
Danske Bank

Mark Seasholes
Arizona State University

We identify long-lived pricing errors through a model in which inattentive investors arrive stochastically to trade. The model's parameters are structurally estimated using daily NYSE market-maker inventories, retail order flows, and prices. The estimated model fits empirical variances, autocorrelations, and cross-autocorrelations among our three data series from daily to monthly frequencies. Pricing errors for the typical NYSE stock have a standard deviation of 3.2 percentage points and a half-life of 6.2 weeks. These pricing errors account for 9.4%, 7.0%, and 4.5% of the respective daily, monthly, and quarterly idiosyncratic return variances. (*JEL* G12, G14)

Received June 11, 2019; editorial decision December 10, 2020 by Editor Stijn Van Nieuwerburgh. Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

How much do observable stock prices deviate from fundamental values? And when they do, how long do these "pricing errors" last? Financial economists

---

We are especially grateful for research help provided by Ariel Lohr and Sunny X. Li. We also thank Hank Bessembinder, Elmar Nijkamp, Paolo Pasquariello, Stefan Ruenzi, Elvira Sojli, and Keke Song for comments as well as seminar participants at HBS, HEC Paris, KAIST (Korea), RMI Singapore, UC Berkeley, UC Santa Cruz, University of Grenoble CERAG, UT Austin, and University of Virginia Darden. We also thank attendees at the NY Fed's Annual Workshop on the Microstructure of Financial Markets, the QMBA Conference at GSM Beijing University, the Liquidity Conference at Erasmus University Rotterdam, the Australasian Finance and Banking Conference, the MFA 2012 conference, and the CICF 2013 Conference. Menkveld gratefully acknowledges Robert Engle and Boyan Jovanovic for sponsoring NYU visiting positions and NWO for a Vici grant. Seasholes acknowledges RGC funding of this project [642509]. Hendershott and Menkveld gratefully acknowledge support from the Norwegian Finance Initiative. Hendershott provides expert witness services to a variety of clients. Hendershott teaches a course for a financial institution that engages in liquidity provision and high-frequency trading activity. Supplementary data can be found on *The Review of Financial Studies* web site.
Send correspondence to Albert J. Menkveld, albertjmenkveld@gmail.com

The Review of Financial Studies 35 (2022) 962–1008

doi:10.1093/rfs/hhab045 Advance Access publication April 15, 2021

have long known that asynchronously arriving (or inattentive) investors could be the root cause of these errors.[1] The pricing errors compensate market makers who supply liquidity by stepping in to match buyers and sellers across time. In a complementary view, the interaction of liquidity supply and pricing errors produces a pattern of predictable return reversals.[2] Short-run reversals are a major focus of the market microstructure literature, while lower-frequency reversals are typically studied in the asset pricing literature. A goal of our paper is to link these two literatures by studying the magnitude of pricing errors for typical New York Stock Exchange (NYSE) stocks at frequencies from a day to a quarter. We find that pricing errors for the typical NYSE stock have a standard deviation of 3.2 percentage points and a half-life of 6.2 weeks. They account for 9.4%, 7.0%, and 4.5% of the respective daily, monthly, and quarterly idiosyncratic return variances.

As alluded to above, a fundamental goal of financial economics is to understand the extent to which a stock's price (or change in a stock's price) reflects a company's fundamental value (or change in value). When an observed price deviates from a company's fundamental value, financial economists seek to understand why and where this "noise" or "pricing error" comes from. Research into noise and inferences about firm values go back at least as far as the 1960s. Fama (1970, 1991) provides concise reviews.[3]

To date, pricing errors have been studied with one of at least three different approaches. First, using only price data, stocks' observed prices can be decomposed into a "fundamental component" and a "transitory component" with sufficient identifying assumptions.[4] However, a purely econometric approach provides few insights into how financial markets work and/or the trade-offs faced by economic agents. Second, infrequent events with large supply shocks can help identify times when pricing errors may appear (see Duffie (2010), for examples). Third, an economic theory model and data on liquidity providers' inventories can be used to identify pricing errors. We follow such an approach in this paper. In our model, different classes of investors have different exposures to private-value shocks. The shocks induce hedging motives to trade and the shocks net to zero across investors (e.g., holdings may change, but do not affect prices if there are no frictions). Our model introduces a friction: some investors continually monitor the stock market, while others are inattentive and arrive infrequently to trade. The presence of the infrequent

[1] Abel, Eberly, and Panageas (2007, 2013) provide microfoundations for how inattention arises endogenously for individual investors. Biais, Hombert, and Weill (2014) model preference uncertainty to capture the inattention of institutional investors.

[2] Examples of such price reversals can be found in Grossman and Miller (1988), Jegadeesh (1990), Lehmann (1990), Campbell, Grossman, and Wang (1993), Llorente et al. (2002), Nagel (2012), and others.

[3] Investor welfare provides further impetus for studying pricing errors. For an example, please see Brennan and Wang (2010). Our pricing error estimates for the magnitude and duration of noise can be used to measure the biases in asset pricing tests shown in Asparouhova, Bessembinder, and Kalcheva (2010, 2013).

[4] For examples, see Roll (1984), Poterba and Summers (1988), Cochrane (1994), and Brennan and Wang (2010).

investors gives rise to the long-lived pricing errors that compensate financial intermediaries for bridging the gap between the needs of frequent and infrequent investors.[5] If infrequent investors arrive over horizons longer than a day, the pricing errors can last for months.

We identify long-lived pricing errors through an economic model. We show how to transform our continuous-time model's theoretical results so that its parameters can be estimated using discretely sampled data. Our data consist of daily NYSE market-maker inventories, retail order flows, and prices. We structurally estimate the pricing errors' magnitude, their duration, as well as deeper economic quantities, such as the risk-bearing capacities of different investor classes. The model yields a flexible form for the pricing errors' data generating process. Inattentive investors can cause long-tailed autocorrelation of market-maker inventories that allows us to more precisely estimate slowly decaying pricing errors.

Using our daily NYSE data on prices, market-maker inventories, and retail trading, we perform maximum likelihood estimation (MLE) to recover the model's underlying parameters. We compare the empirical and model-implied variances and autocorrelations (including cross-autocorrelations among these variables) contemporaneously and with lags ranging from a day to a month. The estimated model matches *all* the relevant dynamic relations, in terms of both signs and magnitudes. This is noteworthy because the model is only a single friction away from a standard asset-pricing model.

The model's rich and long-lasting autocorrelation structure for prices ties together results traditionally in the microstructure literature (daily) with those in the asset pricing literature (monthly and quarterly). Our results show that "noise" is not solely a short-term microstructure effect. Instead, this paper reveals significant pricing errors in monthly data, the data most commonly used in the asset pricing literature. Below, we will expand our discussion of this paper's approach and its contributions.

The model's strength is its simplicity and versatility. For example, our model is invariant to the sampling frequency. Section 2 shows how one can convert the implied model dynamics from continuous time to discrete sampling times, where the latter can span a second, an hour, a day, a month, or a quarter. Our model, therefore, can be used by both monthly asset pricers and submillisecond microstructure researchers. In addition, allowing for multiple classes of slow investors (who operate at different frequencies) is a crucial feature when fitting NYSE price and trading dynamics. In particular, the autocorrelation in daily

[5] Hendershott and Seasholes (2007) and Hendershott and Menkveld (2014) link market-maker inventories to return reversals. Hendershott and Menkveld (2014) uses a discrete-time model with supply and demand shocks arriving independently each period. Their model results in pricing errors following an autoregressive process, specifically, an AR(1). In contrast, our model produces richer pricing error dynamics that better match empirical autocorrelation patterns.

Table 1
Stock return autocorrelations at various frequencies

	Period	Daily	Monthly	Bimonthly
Campbell, Lo, and MacKinlay	1962–1994		−0.03	−0.04
Our data	1999–2005	−0.02	−0.04	−0.08
Model implied	1999–2005	−0.01	−0.04	−0.05

This table presents first-order autocorrelations of individual stock returns. It illustrates that longer period returns can have *more* negative first-order autocorrelations. The Campbell, Lo, and MacKinlay (1997, p. 73, table 2.7) results are based on a mapping from their variance ratios to first-order autocorrelations (see their eq. (2.8.1) on p. 69). Their results are based on individual returns for 411 U.S. stocks. Our data are more recent and based on idiosyncratic returns for 689 U.S. stocks. The model-implied autocorrelations are based on estimates presented in Section 3.

idiosyncratic returns[6] decays too slowly to be explained using only a single class of slow investors. We find a good fit using three classes of slow investors: one with investors who arrive daily (on average), a second with investors who arrive monthly, and a third with investors who arrive quarterly.

The monthly and quarterly inattentive investors lead to the slowly decaying pricing errors found in the autocorrelations of NYSE returns. The presence of these classes is also the main reason pricing errors are sizeable. We estimate that prices deviate from fundamental values by 3.2 percentage points with a half-life of 6.2 weeks.

The slow decay in pricing errors can further explain a (perhaps) puzzling empirical feature of NYSE data: first-order return autocorrelations can become more negative when sampled at lower frequencies. Table 1 illustrates this puzzle for both a classic and a modern sample of U.S. equities. Campbell, Lo, and MacKinlay (1997) find that stock-specific returns are more negatively autocorrelated at a bimonthly frequency than at a monthly frequency for their 1962–1994 sample. In our 1999–2005 sample, we find a similar pattern when comparing daily, monthly, and bimonthly returns. The table further shows that our model can produce such a pattern. The model-implied autocorrelations become increasingly more negative as one moves from daily to monthly and then to bimonthly returns. The intuition for why such patterns can occur is presented in Appendix A. Additional details are given in Internet Appendix C. If pricing-error persistence is large, such errors will wash out in returns (at high frequencies) as the first-order autocorrelation will tend to zero. At low frequencies, the errors will decay enough to cause a negative first-order autocorrelation. This insight should caution researchers not to conclude that prices are “efficient” when seeing negligible first-order autocorrelation in returns sampled at high frequencies. Our results suggest it is very difficult to separate pricing errors from fundamental values using only observable prices in a finite sample. Trading data can help avoid such difficulties.

[6] All returns in the paper are idiosyncratic. Hereafter, to ease exposition, we typically refer to them simply as “returns.”

Our structural model estimation yields novel insights in four broad areas. First, the model requires a range of slow investor classes in order to achieve a reasonable fit: we use daily slow investors, monthly slow investors, and quarterly slow investors. These classes feature both slow institutional and retail investors. Institutions are more prevalent at all three frequencies.[7] While retail investors are a small part of the market, they make up a relatively larger part of the monthly and quarterly slow investors.[8] These observations are based on our estimates of the total masses of private-value shocks, referred to as "risk masses." This term emphasizes that it is the product of the mass of investors times the size of the per-investor private-value shock. In other words, while the model is unable to identify how many investors are in each class or the hedge shocks they experience, we are able to identify the product of the two.

Second, the model allows for a decomposition of the pricing error variance. The standard deviation of the various components are 0.097%, 1.575%, and 2.548% due to the respective daily, monthly, and quarterly slow investors and 1.106% due to a component shared across all the investor arrival classes. In addition to producing results for our sample of NYSE stocks, we also produce them for three, size-based subsamples of stocks (large, medium, and small stocks).

Third, the model quantifies the price impact of institutional trading. A \$192 million shock to fast institutions' target portfolios leads to a pricing error of only 1.3%. This indicates that there is substantial risk-bearing capacity at the time of the initial shock (both in terms of market making and in terms of institutions ability to patiently trade.)

Fourth, we are able to carry out a counterfactual analysis. We find pricing errors explain 9.4% of daily return variance. We then vary the risk aversion of the fast investors. Also, we assume slow institutions start reacting faster (perhaps because of technology improvements). Not surprisingly, having twice as many risk-tolerant fast investors (or having quarterly and monthly slow institutions become daily slow investors) dramatically reduce the pricing errors' fraction of daily return variance (from 9.4% to 2.5% and 0.9%, respectively).

Our paper is closely tied to a literature that started with Grossman and Miller (1988) in which market makers smooth nonsynchronous trading demands due to inattentive investors. Recent inattention papers, such as Duffie (2010) and Bogousslavsky (2016), include attention heterogeneity that increases the need

[7] Since we have market-maker inventories and retail trades, institutional trades are defined by a market-clearing constraint. Lakonishok, Shleifer, and Vishny (1992) measure the size of the institutional imbalance and its relation to current price movements. Other papers study interactions of institutional and retail trading (see papers such as Nofsinger and Sias (1999) and Griffin, Harris, and Topaloglu (2003)). Our paper speaks to both literatures. We can measure the magnitude of pricing errors and relate it to buy-sell imbalances of any of our investor classes.

[8] Stock trading by retail investors is well studied, and the most relevant paper is Kaniel, Saar, and Titman (2008). The authors show that net trades by retail investors this week are positively related to returns the following week. We confirm the earlier results and add new economic insights based on the inattention friction. Not surprisingly, we estimate retail investors to be a small fraction of slow investors. We further find that, in relative terms, they are a larger part of quarterly and monthly slow investors than of daily slow investors.

for intertemporal smoothing.[9] Bogousslavsky (2016) shows that the inattentive investors can explain regularities in stock return autocorrelation patterns. Our model differs from these papers in a number of key ways. First, our model has market makers, attentive (fast) investors, *and* multiple classes of inattentive (slow) investors. Importantly, our inattentive investors arrive stochastically. This feature keeps the dimensionality of the state space small enough to make structural estimation feasible and thus identification of pricing errors possible (a detailed argument can be found at the end of Section 1.2). Our closed-form solutions also allow us to decompose the pricing errors into easily understood economic quantities.

Duffie (2010) discusses a number of empirical examples where pricing errors are found by identifying liquidity demand shocks. Koijen and Yogo (2019) provide systematic evidence on liquidity demand by using changes in 13F (holdings) data and changes in prices to estimate the latent demand of institutions at a quarterly frequency. Under the assumption that this latent demand is mean reverting, Koijen and Yogo (2019) find that institutions can cause long-lived price pressure in the cross-section of stock returns. This finding complements our stock-level findings that pricing errors (in the time series dimension) are identified by liquidity supply (market-maker inventories). Cella, Ellul, and Giannetti (2013) examine how the institutional investors' average holding periods across stocks relate to pricing errors during marketwide negative shocks. They find that stocks held more by short-horizon investors experience larger price drops and subsequent reversals. If short-horizon holding periods correspond to more frequent rebalancing needs, their results are consistent with our model and empirical results that larger hedging shocks lead to larger pricing errors.

Pricing errors arise in studies of bond and currency markets as well. Bao, Pan, and Wang (2011) assume prices follow a random walk and estimate illiquidity as the negative covariance of high-frequency and daily price changes. Hu, Pan, and Wang (2013) construct a marketwide noise measure by backing out the implied yield curve from the daily cross-section of bonds and bills. Bacchetta and van Wincoop (2010) calibrate a two-country model with infrequent portfolio rebalancing. Their results of a forward discount bias mirror empirical findings that have long puzzled economists.

1. Asset Pricing Model with Limited Attention

Our theoretical model is recursive in nature, assumes that all investors are price-takers, and runs in continuous time. The model's core distinguishing feature is the inclusion of multiple classes of inattentive investors who operate

[9] Chien, Cole, and Lustig (2012) explore how inattention in the form of intermittent rebalancing increases the volatility of the market price of risk. Crouzet, Dew-Becker, and Nathanson (2019) and Weller (2018) examine how short-term investors affect the incentives of long-term investors to acquire information about firms.

at different frequencies. Such inattention is the only friction in the model (i.e., information is symmetric and agents are zero-mass price-takers).[10] Our model includes private-value shocks that investors experience and which offset one another. Therefore, in the absence of the inattention friction, trade is purely reallocational, does not require intermediation, and does not affect prices (i.e., no pricing errors). However, if at least one investor class is inattentive the model can generate nontrivial price and trade patterns.

Intuition for the channels that generate our trading and return patterns can be obtained by considering an example subsumed by our model. Consider investors who might experience private-value shocks for a single asset. Let part of the investor mass experience no such shocks and be perfectly attentive, meaning they are continuously present in the market and ready to trade. These investors will endogenously become market makers. Divide the remaining investor mass in half and let the private-value shocks that one-half experiences be offset by the shocks that the other half experiences. In other words, the target-holding changes for the asset sum to zero. Let one-half be perfectly attentive (fast), like the market makers, and the other half be inattentive and arrive to trade with (stochastic) delays.

Now consider that the attentive investors receive a negative private-value shock. As the inattentive or slow investors are not all there at the time of the shock, prices temporarily experience downward pressure to clear the market. This negative pricing error attracts market makers who purchase the securities that the fast investors want to sell.[11] It also induces these fast investors to reduce their current liquidity demands and spread these demands over time (i.e., the optimal trading strategy calls for "parceling out" trades).[12] Both the fast investors and the market makers will sell to slow investors once the latter investor class arrives at the market in the future, and as a result, the pricing error will subside. The magnitude of the shocks, the relative sizes of the different investor classes, and the inattention frequency of the slow investors together determine the magnitude and duration of the pricing errors.

1.1 Model primitives

Time is continuous, indexed by t, and runs forever. Setting the model up in continuous time yields closed-form expressions that serve three purposes. First,

10 Having only one friction (inattention) both clarifies the channels at work in the model and disciplines the data-fitting exercise.

11 Modeling the inventory control choices of market makers is highlighted by both Madhavan and Smidt (1993) and Hendershott and Menkveld (2014). Our paper treats market makers as competitive price takers and does not allow them to trade strategically. Such an assumption is helpful for obtaining closed-form solutions. At short horizons the NYSE market makers have information and positional advantages that likely enable them to behave strategically. These advantages diminish at lower frequencies, making NYSE market makers compete with hedge funds and other investors to provide liquidity at longer horizons.

12 This links our paper to the optimal execution literature (see Bertsimas and Lo (1998) and Almgren and Chriss (2001), for examples). In our model, both the market makers and fast investors solve for optimal trading strategies, except with different goals, leading to endogenous pricing errors.

the setup creates transparent relationships between the model's deep parameters and the economic variables of interest. This transparency facilitates economic insights. Second, the closed-form expressions make structural estimation feasible. Third, our model becomes invariant to the sampling frequency. Section 2 shows how to convert the implied model dynamics from $\mathrm{d}t$ to Δt where the latter can span a second, an hour, a day, or a month. Our model, therefore, can be used by monthly asset pricers as well as submillisecond microstructure researchers. Appendix B provides a summary of the notation used in our model.

Assets. There are two assets in the economy. First, there is a risky asset in zero net supply that pays dividends over any interval $(t, t+\mathrm{d}t]$, with B being a Brownian motion:

$$\mathrm{d}D_t = \sigma_w \mathrm{d}B_t. \tag{1}$$

The dividend process, having an expected value of zero, implies that the asset's fundamental value is zero. However, this dividend process is consistent with modeling a pricing error that fluctuates around zero. Of course, adding a positive expected dividend would cause the asset's expected price to be above zero. Given that this paper's empirical focus is on pricing errors and price changes, it becomes convenient to center the dividend dynamics around zero. Second, there is a risk-free asset with an exogenously given rate of return $r > 0$. The risk-free asset is in perfectly elastic supply ensuring a constant payoff.

Investors. There are $N+2$ classes of investors: fast institutions (indexed by F), market makers (indexed by M), and $N \in \mathbb{N}$ classes of slow investors (indexed by $i = 1, \ldots, N$). Let $\mathbf{N} := \{1, \ldots, N\}$ denote all classes of slow investors. We index all of the $N+2$ classes with $j \in \{F, M\} \cup \mathbf{N}$. There is a continuum of agents in each of the $N+2$ classes. The masses of the investor classes are m_F, m_M, m_1, …, m_N, respectively.

The slow investors are inattentive and only trade the risky asset infrequently. Concretely, a slow investor belonging to class i trades the risky asset at the jump times of a Poisson process. The jump intensity of this Poisson process is λ_i and the Poisson processes are independent across investors (even within a class).[13] For convenience, we define Λ to be the diagonal matrix whose entries are the attention intensities of the slow investors:

$$\Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_N). \tag{2}$$

[13] A more general setting would allow for a correlation across the inattention processes. For example, one could add common shocks that would bring all inattentive investors to the risky asset market at the same time. In such a case, the price jumps toward its efficient level (i.e., to zero in our setting). Furthermore, even when not all investors are paying attention, the possibility of this abrupt convergence induces bolder bets against inefficient prices. Overall, making the attention processes correlated across agents attenuates the effect of inattention on prices, but does not eliminate the qualitative results.

Preferences. All investors are risk neutral but suffer a quadratic utility loss when their holdings of the risky asset deviate from a certain target. This target is moving over time and shared by investors within a given class (more details can be found in a few paragraphs). Concretely, at time t, an investor i of class j chooses his policies to maximize

$$\sup_{C,\pi} \mathrm{E}_t\left[\int_t^{\infty} e^{-r(u-t)}\left(\mathrm{d}C_u - \frac{r\gamma_j\sigma_w^2}{2}\left(T_{j,u}-\pi_{i,u}\right)^2 \mathrm{d}u\right)\right], \tag{3}$$

where C_u is the cumulative consumption of the investor up to time u, $T_{j,u}$ is the target portfolio for class j at time u, $\pi_{i,u}$ denotes his actual risky asset holdings at time u, and $\gamma_j > 0$ is a risk-aversion parameter that determines the utility loss per unit of differential between target and actual holdings.

We interpret preferences as specified in (3) as follows: a class j investor wants to hedge some risky exposure and can do so perfectly by holding $T_{j,t}$ shares of the risky asset. If the expected excess return on this asset is not currently zero, then a speculative position in the risky asset will increase the investor's expected wealth and consumption. The optimal portfolio balances hedging benefits and speculative profits. The quasi-linear preferences of (3) are similar to those in Biais (1993), Duffie, Gârleanu, and Pedersen (2007), Gârleanu (2009), Lagos and Rocheteau (2009), and Afonso and Lagos (2015).

Target portfolios. An N-dimensional Brownian motion, Z, drives the slow investors' target portfolios (the innovations to Z_t are also referred to as "hedge shocks" in this paper). Concretely, the target portfolio vector that comprises all slow investor classes is shown below. The first term in (4) is the volatility of the target shocks experienced by each of the N slow investor classes:

$$T_{\mathbf{N},t} := diag(\sigma_1,\ldots,\sigma_N)\,Z_t \in \mathbb{R}^N. \tag{4}$$

The target portfolio of the market makers is zero at all times and is shown in (5). This definition is consistent with market makers only trading to facilitate risk sharing among the other market participants:

$$T_{M,t} := 0. \tag{5}$$

Finally, the (scalar) target portfolio of the fast institutions is shown in (6), where $\mathbf{1}_{(k\times l)}$ is a $k\times l$ matrix of ones:[14]

$$T_{F,t} := -\frac{1}{m_F}\mathbf{1}_{(1\times N)}diag(m_1,\ldots,m_N)\,T_{\mathbf{N},t}, \tag{6}$$

[14] In our model, we abstract away from target shocks affecting the risky asset's fundamental value. We follow Lo, Mamaysky, and Wang (2004) in assuming that the fast institutions' target portfolio is equal and opposite to a weighted sum of the slow investors' targets. Additional details on the daily target portfolio changes are in Internet Appendix D.

With the target portfolios defined in (4), (5), and (6), the sum of the target holdings in the risky asset is zero at all times:

$$\sum_{j\in\{F,M\}\cup\mathbf{N}} m_j T_{j,t} = 0. \tag{7}$$

If all investors are attentive at all times, then all investors will always hold their target portfolios, and there is no reason for the price to differ from fundamental value (i.e., what is often referred to as the "permanent component of price" is zero in our setting).

The Brownian motions in our paper are allowed to be correlated (i.e., the B_t that drives the dividend process and the Z_t's that drive the target portfolios). Specifically, a correlation of ρ links the âŁ˜âŁ˜with dividend" price/return dynamics and a shared target portfolio shock to all investors:

$$\mathrm{Corr}(\mathrm{d}B_t, \mathrm{d}Z_t) = \rho \cdot \mathbf{1}_{(N\times 1)}. \tag{8}$$

Equation (8) is a reduced-form way to model correlation between the permanent component of price and shocks faced by slow/fast investors.[15] Such a correlation could arise from target portfolio shocks being imbalanced between fast and slow investors such that the sum of the shocks is nonzero.[16] In this case, the permanent component of price will adjust so that market clearing occurs at the long-run/permanent price where fast and slow investors' target portfolios (conditional on the new equilibrium price) sum to zero. Appendix C illustrates how imbalanced shocks can yield a permanent price process that is equivalent to the with-dividend permanent price process used here. More broadly, Appendix C illustrates how imbalanced shocks can result in a correlation of ρ between the balanced shock process and returns.

The gap process (state variable). Finally, it is useful to define a "gap process" or G_t across all classes of slow investors. This process keeps track of the gaps between the target and actual portfolios and is summed across all slow investors in the N different classes. More precisely,

$$G_t := diag(m_1, \dots, m_N)\big(T_{\mathbf{N},t} - A_{\mathbf{N},t}\big) \in \mathbb{R}^N, \tag{9}$$

where entry i of $A_{\mathbf{N},t} \in \mathbb{R}^N$ contains the actual holdings of all investors in class i:

$$A_{i,t} := \int_{u\in m_i} \pi_{u,t}\,\mathrm{d}u. \tag{10}$$

This gap process is the state variable on which all the model's dynamics depend. Defining the gap process at an investor-class level benefits from the independent

15 By "with-dividend permanent price," we mean the accumulation of all dividends up to the current time, $P_T = \int_0^T \mathrm{d}D_t$.

16 Such a correlation could also arise from information-based trading. For example, ρ would be negative if fast investors trade on information in addition to their target portfolio shocks.

arrivals of the investors within the class. A "law of large numbers" result holds and consequently the gap process is an Ornstein-Uhlenbeck (OU) process [an AR(1) process in continuous time].

The OU (gap) process has economic appeal as it essentially captures the order imbalance relative to a first-best (i.e., the case when all investors are fully attentive). Because the gap process represents an imbalance, market-clearing prices and their dynamics depend on it. This dependence will become clear in the next subsection where we present equilibrium results. We will also show that *changes* in the gap process relate to market-maker inventories and slow-investor flows.

1.2 Equilibrium

To ensure that the model's full dynamics become available in closed form, we assume slow investors are infinitely risk averse (i.e., $\gamma_j = \infty, \forall j \in \mathbf{N}$).[17] This is a technical assumption that removes speculation by slow classes.[18] This makes inattentive investors act like liquidity traders in many models. Such traders do not act strategically nor condition their trading on price.

Our main equilibrium result is based on standard definitions that feature individual optimality, market clearing, and rational expectations (see Appendix D.3. for this definition and a proof of the following proposition). Our approach to solving for an equilibrium can be categorized as "guess and verify." We first solve the individual problems for all agents assuming a price process for the risky asset. Then, given these solutions, we show that the assumed price process is the result of market clearing. Appendix D.1. provides a more detailed description.

We can write the gap process as follows:

$$\mathrm{d}G_t = -\Lambda G_t \mathrm{d}t + \mathrm{diag}(\mu_1, \ldots, \mu_N)\mathrm{d}Z_t, \tag{11}$$

where $\mu_j := m_j \sigma_j$ is the total *risk mass* of investors in class j. Appendix D.1. discusses the three key assumptions and/or guesses: (1) the gap process follows an Ornstein-Uhlenbeck ("OU") process; (2) the pricing errors are linear in the gap process; and (3) the gap process is public information.[19]

---

[17] Prior versions of this paper presented an extended version of the Duffie (2010) model. That model allows for market makers and multiple classes of slow investors who trade strategically. The model with slow investors trading strategically results in predictions qualitatively similar to those in the current draft: all of the previous model's moments have the same sign as those in the continuous-time model with nonstrategic slow investors. However, our extended Duffie (2010) model does not have closed-form solutions and is not suitable for structural estimation. Finally, please note that myopia, which limits strategic behavior, is used in models such as Nagel (2012) and facilitates obtaining closed-form solutions.

[18] Investors who trade monthly are more likely to trade to exactly their target portfolio because the cost of speculative trading (quadratic loss) is greater the longer the duration between the investor's trades.

[19] As agents are risk neutral in terms of consumption with a time preference parameter equal to the interest rate, any policy in which consumption eventually takes place is equally good. Therefore, no consumption policy is reported. Note that delaying consumption forever is not optimal.

An OU process for the gap vector in (11) has intuitive appeal as mentioned earlier. We see that class j's gap decays smoothly with intensity λ_j (an element in Λ). The independent arrivals of investors generate the smoothness. The size of gap shocks scales with the mass of investors in this class since μ_j is the product of m_j and the size of an individual-investor shock (σ_j).

The equilibrium price process and optimal holdings of all agents are available in closed form (with a proof in Appendix D).

Proposition 1 (Equilibrium Price Process and Holdings). An equilibrium exists and is unique among equilibria where the gap process follows an Ornstein-Uhlenbeck, pricing errors are linear in the gap process, and the gap process is public information. The first expression is for the price process. The final three expressions are for holdings:

- The equilibrium price of the risky asset is given below where $p \in \mathbb{R}^N$ and I_N is the identity matrix in $\mathbb{R}^{N \times N}$:

$$P_t = -p^\top G_t \qquad \text{with} \qquad p^\top = \frac{\sigma_w^2}{\frac{m_F}{r\gamma_F} + \frac{m_M}{r\gamma_M}} \mathbf{1}_{(1\times N)} (r I_N + \Lambda)^{-1}. \tag{12}$$

- A market maker holds $\pi_{M,t}$ shares of the risky asset:

$$\pi_{M,t} = \frac{1}{r\gamma_M \sigma_w^2} \left[\frac{1}{\mathrm{d}t} \mathrm{E}_t(\mathrm{d}P_t) - r P_t \right] = \frac{1}{r\gamma_M \sigma_w^2} \left[p^\top (r I_N + \Lambda) G_t \right]. \tag{13}$$

- A fast institution holds $\pi_{F,t}$ shares of the risky asset:

$$\pi_{F,t} = T_{F,t} + \frac{1}{r\gamma_F \sigma_w^2} \left[p^\top (r I_N + \Lambda) G_t \right]. \tag{14}$$

- A slow investor of class j who arrives at the market at time t holds $\pi_{j,t}$ shares:

$$\pi_{j,t} = T_{j,t}. \tag{15}$$

Proposition 1 leads to the following observations. The equilibrium price process determines the dynamics of the trading policies of the market makers and fast institutions (see the row vector of weights, $p^\top$, in (12)).

Second, the price impact row vector $p^\top$ that translates portfolio-holding gaps to pricing errors yields several insights. Higher fundamental risk (σ_w) or lower risk absorption capacity of fast investors increase the price impact. This is not surprising. What is not as obvious is that a one-unit larger gap for class j investors commands a price impact that is inversely proportional to the arrival intensity of investors plus the risk-free rate. Investors in our model require a larger compensation for speculating against slower investors. This result is intuitive, as fast investors are stuck with a position for longer.[20]

Third, Proposition 1 shows how hedging and/or speculative motives define the various optimal portfolios. Starting with the market maker's holdings in (13), note that the RHS term loads positively on a weighted sum of the gap process with weights proportional to the row vector $p^\top$. In (12), the pricing error loads negatively on this same weighted sum. By holding more of an asset with a negative pricing error (that will mean revert to zero), the market maker makes a speculative profit in expectation. Further, note that the speculative motive is larger when he is less risk averse (γ_M) or when the asset has less fundamental risk (σ_w). The fast institution's portfolio in (14) features both hedging and speculative motives additively. The first RHS term involves his target portfolio and therefore represents hedging. The second RHS term is the speculative motive. Note that the slow investor's portfolio in (15) only features a hedging motive as, by assumption, this investor classes does not engage in speculation.

Fourth, expressions for the optimal holdings yield an interesting observation. As noted when discussing Proposition 1, more fundamental risk reduces the speculative positions of fast investors (i.e., fast institutions and market makers), *all else equal*. In equilibrium, however, the same logic does not follow, and speculative positions are invariant to fundamental risk. The compensation for bearing fundamental risk increases in equilibrium to the point that fast investors willingly take it on (i.e., note that the σ_w^2 in (14) cancels against σ_w^2 in (12)). This result is best understood by market clearing. The risky positions have to be held by the fast investors as they are the only ones with positive risk-bearing capacity (i.e., $\gamma_F, \gamma_M < \infty, \gamma_j = \infty$).

Finally, note that the dimensionality of the state variable G_t depends on the number of slow-investor classes N and can therefore be kept small during estimation (e.g., $N=6$ in Section 3.2). Yet, pricing errors can stretch across long horizons depending on how inattentive the slowest investor is. This is an important feature of our model as it makes structural estimation possible. One can compare our stochastic-arrivals setup to the setup found in a model such as Duffie (2010) which features infrequent but deterministic arrivals. Such a model needs a state space with dimensionality equal to the frequency of the slowest investors. If one wants to generate monthly effects using daily data, this requires a state-space of dimensionality 21. An additional benefit of our model is that it yields analytic expressions for any dimensionality, while Duffie's model generally does not.

[20] The larger premium for lower interest rates and, at the same time, less discounting—see the preferences in (3)—is more challenging to explain. It appears that temporarily tying up capital in speculative positions is more expensive in our economy.

2. Model-Implied Discrete-Time Dynamics

This section translates the continuous-time model to a version that makes estimation possible for discrete-time data sampled in Δt periods. The section first derives the model dynamics to provide expressions for variances, covariances, and autocorrelations. Appendix B provides a notational summary of the parameters used in estimation. In Section 3, we use maximum likelihood to estimate the model's parameters using NYSE data. Our data's sampling period is one day. To keep the structural estimation numerically tractable, we consider three classes of limited-attention (slow) investors:

- Class d investors who, on average, arrive at the market once a *d*ay,
- Class m investors who, on average, arrive once a *m*onth, and
- Class q investors who, on average, arrive once a *q*uarter.

As our data are daily, we pick one class (d) to match this frequency. We then add a slower class (m) and a much slower class (q).[21] The slow investors arrive at the market with Poisson intensities such that average durations are once a day (for d), once a month (for m), and once a quarter (for q). For each of the slow investor classes, we further categorize into two subclasses. Slow investors are either institutional ("i") or retail ("r"). The reason for this further categorization is that we have retail-flow data.[22] The investor-class subscripts are thus $\{d,m,q\}\times\{i,r\}$.

The matrix with Poisson intensities is given by

$$\begin{aligned} \Lambda_j &= \operatorname{diag}\left(\lambda_{dj},\lambda_{mj},\lambda_{qj}\right)=\operatorname{diag}\left(1,\frac{1}{21},\frac{1}{63}\right), \quad j\in\{i,r\} \qquad (16) \\ \Lambda &= \begin{bmatrix} \Lambda_i & 0 \\ 0 & \Lambda_r \end{bmatrix} \in \mathbb{R}^6\times\mathbb{R}^6. \end{aligned}$$

The gap processes for the 3×2 classes of slow investors become[23]

$$\begin{aligned} G_{j,t} &= \left(G_{dj,t} \quad G_{mj,t} \quad G_{qj,t}\right)^\top \in\mathbb{R}^3, \quad j\in\{i,r\} \qquad (17) \\ G_t &= \begin{bmatrix} G_{i,t} \\ G_{r,t} \end{bmatrix} \in\mathbb{R}^6. \end{aligned}$$

[21] Internet Appendix F shows that adding an intermediate frequency, such as weekly, does not provide additional insights. The model simply puts no weight on the risk masses of the weekly investor classes.

[22] We refer to individuals as "retail," so we can use different single-letter subscripts for institutions and individuals.

[23] The model estimation procedure is based on maximum likelihood estimation (MLE). In principle, there is no bound to the number of classes and/or frequencies a researcher could study. However, one runs into issues with dimensionality of the parameter space and troubles inverting key matrices. We have three classes of slow investors and have chosen natural frequencies that match our data frequency and existing empirical work. Internet Appendix F shows that adding an intermediate frequency of investors (i.e., weekly) does not provide additional insights.

The gap processes are associated with investors that have risk masses (i.e., $\mu = m\sigma$):

$$\begin{aligned} \mu_j &= \begin{pmatrix}\mu_{dj} & \mu_{mj} & \mu_{qj}\end{pmatrix}^\top \in \mathbb{R}^3, \quad j \in \{i, r\} \\ \mu &= \begin{bmatrix}\mu_i \\ \mu_r\end{bmatrix} \in \mathbb{R}^6. \end{aligned} \tag{18}$$

Discrete-time dynamics. The discrete-time dynamics for the full model can now be written down explicitly. First, we stack all the model variables in the following vector:

$$Y_t = \begin{pmatrix} G_t^\top & MMInv_t & RetFlow_t & Return_t \end{pmatrix}^\top \in \mathbb{R}^9, \tag{19}$$

where G_t is defined above, $MMInv_t, RetFlow_t, Return_t \in \mathbb{R}$ are the end-of-period market-maker inventories, per-period retail flows, and per-period returns, respectively (where period t runs from time $t-1$ to time t). The model-implied dynamics are

$$Y_t = V Y_{t-\Delta t} + W \varepsilon_t. \tag{20}$$

The dynamics in (20) imply a vector autoregression (VAR) in which the coefficient matrix V incorporates the autoregressive component and the coefficient matrix W maps the shocks into the model's variables. The VAR cannot be estimated directly because the elements of G_t are not directly observable in the data.

Writing out the model's discrete-time dynamics shows how the different model parameters affect the model's dynamics and allows for structural estimation. The coefficient matrix V (with row and column dimensions in light gray along the axes) is

$$V = \begin{array}{cc} & \begin{array}{ccccc} 3 & 3 & 1 & 1 & 1 \end{array} \\ \begin{array}{c} 3 \\ 3 \\ 1 \\ 1 \\ 1 \end{array} & \begin{bmatrix} e^{-\Lambda_i \Delta t} & 0 & 0 & 0 & 0 \\ 0 & e^{-\Lambda_r \Delta t} & 0 & 0 & 0 \\ \beta_M \mathbf{1}_{(1\times 3)} e^{-\Lambda_i \Delta t} & \beta_M \mathbf{1}_{(1\times 3)} e^{-\Lambda_r \Delta t} & 0 & 0 & 0 \\ 0 & \mathbf{1}_{(1\times 3)} \left(I_3 - e^{-\Lambda_r \Delta t}\right) & 0 & 0 & 0 \\ \beta_w \mathbf{1}_{(1\times 3)} A_i & \beta_w \mathbf{1}_{(1\times 3)} A_r & 0 & 0 & 0 \end{bmatrix} \end{array} \in \mathbb{R}^{9\times 9}, \tag{21}$$

where I_n is the identity matrix of size n, $A_j = \left(r I_3 + \Lambda_j\right)^{-1}\left(I_3 - e^{-\Lambda_j \Delta t}\right)$ with $j \in \{i, r\}$, and the betas are defined as follows:

$$\beta_w = \frac{\sigma_w}{\frac{m_F}{r\gamma_F} + \frac{m_M}{r\gamma_M}} \quad \text{and} \quad \beta_M = \frac{\frac{m_M}{r\gamma_M}}{\frac{m_F}{r\gamma_F} + \frac{m_M}{r\gamma_M}}. \tag{22}$$

These two betas capture (ratios of) deep economic parameters from our model and are discussed further in the next two paragraphs. We refer to $\frac{m_j}{r\gamma_j}$ as the "risk-aversion adjusted mass of investor class-j." In our model, two of the investor classes (M and F) are present at the time of a shock. Each class conditions its behavior on the price impact it may have. Thus, the risk-aversion-adjusted masses of both appear in (22).

The first beta in (22), β_w, is the ratio of the asset's fundamental risk (σ_w) to the sum of the risk-aversion-adjusted masses of the market makers and fast institutions. In other words, it is the magnitude of a typical shock divided by how many investors (adjusted) are immediately around to trade. Note also that $\beta_w \sigma_w$ is also the first term in the factor ($p^\top$) that scales the gap process as shown in (12). If β_w is zero, then investors' inattention does not affect prices. β_w can be zero if the dividend process has zero variance, if the market makers are risk neutral, or if the fast investors are risk neutral.

The second beta in (22) is β_M, which captures the market makers' fraction of risk-aversion-adjusted mass available to trade at the time of the shock. This is important because the larger market makers are relative to the trading needs of the fast institutions, the more markets will accommodate the fast institutions' immediate trading needs. Hence, β_M plays a significant role in the market-maker inventory dynamics, while β_w plays a significant role in the price dynamics. Both betas are proportional to the gap processes, but with different sensitivities.

The various elements of V are intuitive. The first three rows capture what (in expectation) at the end of the period is left of start-of-period inefficient holdings of the three classes of inattentive institutions. The change is due to in-period arrivals of some of these institutions. The same goes for the second set of three rows, which correspond to retail investors. The seventh row sums these residual inefficient holdings across all slow investors to identify how they contribute to end-of-period market-maker inventory. The eighth row picks up the flow of in-period retail investor arrivals by subtracting their end-of-period inefficient holdings from their start-of-period inefficient holdings (i.e., $I_3 - e^{-\Lambda_r \Delta t}$). The ninth row also picks up this flow for inattentive institutions and retail investors to capture how much of the pricing error disappeared due to in-period arrivals.

2.1 Intuition

Appendix E gives a description of the dynamics of the full model with nine variables. This includes the various elements of W and the variance and covariance of the shocks, ε_t. To gain intuition about how the VAR and our data identify model parameters, we simplify (19) from $\mathbb{R}^9$ to $\mathbb{R}^3$ by focusing only on one gap process (one class of slow investors), market-maker inventories, and returns. In addition, we set $\rho = 0$ for the time being.[24] Thus, the VAR in (20) has a reduced form using $Y_t = \begin{pmatrix} G_t & MMInv_t & Return_t \end{pmatrix}^\top$ in this example. The V matrix in (21) is similarly reduced for this simplified version of the model by considering only rows and columns numbered 1, 7, and 9.

[24] The full system in Appendix E is used in the empirical estimation, and it allows for additional classes of slow investors, retail order flows, and $\rho \neq 0$.

The lag k (>0) autocovariance matrix of the reduced Y_t vector is shown below. We omit the first row and column that correspond to the unobservable gap process and focus on the lower-right 2×2 elements that relate to $MMInv_t$ and $Return_t$:

$$Cov(Y_t, Y_{t-k})$$
$$= e^{-k\lambda\Delta t}\cdot\frac{\mu^2}{2\lambda}\cdot\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \beta_M^2 & -\beta_M\beta_w\left(\frac{1-e^{-\lambda\Delta t}}{r+\lambda}\right) \\ \cdot & \beta_M\beta_w\left(e^{+\lambda\Delta t}\right)\frac{1-e^{-\lambda\Delta t}}{r+\lambda} & -\beta_w^2\left(e^{+\lambda\Delta t}\right)\left(\frac{1-e^{-\lambda\Delta t}}{r+\lambda}\right)^2 \end{bmatrix}, \tag{23}$$

where both β_w and β_M are defined in (22) earlier. The structure of (23) indicates that the variance of returns and market-maker inventories should help enable identification of β_w and β_M. The entire matrix is multiplied by an exponential decay factor in λ and the constant $\frac{\mu^2}{2\lambda}$. Because μ does not appear anywhere else we must use more than just the variances and covariances of returns and market-maker inventories to identify μ. This motivates our use of autocorrelations and cross-autocorrelations.

From the above autocovariance matrix we can calculate covariances, cross-autocovariances, variances, correlations, and cross-autocorrelations. We begin with market-maker inventories which have a variance of

$$Var(MMinv_t) = \frac{\mu^2}{2\lambda}\cdot\beta_M^2. \tag{24}$$

Once μ and λ are identified the variance of market-maker inventories identifies β_M. The autocovariance of market-maker inventories is

$$Cov(MMinv_t, MMinv_{t-k}) \;=\; e^{-k\lambda\Delta t}\cdot\frac{\mu^2}{2\lambda}\cdot\beta_M^2.$$

Combining these into the autocorrelation of market-maker inventories yields

$$Corr(MMinv_t, MMinv_{t-k}) \;=\; e^{-k\lambda\Delta t}. \tag{25}$$

The market-maker inventories follow an OU process which leads to their decay following an AR(1) process. Hence, the dynamics of market-maker inventories identifies the arrival intensity of the slow investors (the λ parameter). Note that while there is only one λ in this simplified version of the model, we consider multiple classes of slow investors in our empirical analysis. In the more general setting, market-maker inventories continue to help identify the arrival intensities of the different classes of slow investors. Additional classes of slow investors cause market-maker inventories to follow a multidimensional OU process. This requires multiple values of k to be used to help identify the different arrival intensities of the different classes of slow investors. The intuition from the simple model is that if the estimated multiple-slow-investor-class model fits the autocorrelation of market-maker inventories in the data, then the number and arrival intensities of the classes of slow investors is well identified.

Next, we turn to returns, which have a variance of

$$Var(Ret_t) \quad = \quad \frac{\mu^2}{\lambda} \cdot \beta_w^2 \cdot \frac{\left(1-e^{-\lambda\Delta t}\right)}{(r+\lambda)^2} + \sigma_w^2 \Delta t. \tag{26}$$

The first term of this variance is increasing in μ, while both terms are increasing in σ_w because β_w is proportional to σ_w (see (22)). In addition, the variance of returns can be decomposed to reflect both the slow reduction of legacy pricing errors and new shocks that arrive:[25]

$$Var(Ret_t) = \overbrace{\frac{\mu^2}{2\lambda} \cdot \beta_w^2 \cdot \frac{\left(1-e^{-\lambda\Delta t}\right)^2}{(r+\lambda)^2}}^{\text{Legacy error reduction}} + \overbrace{\frac{\mu^2}{2\lambda} \cdot \beta_w^2 \cdot \frac{\left(1-e^{-2\lambda\Delta t}\right)}{(r+\lambda)^2} + \sigma_w^2 \Delta t}^{\text{New shocks}}.$$

The autocovariance of returns displays a similar AR(1) decay as does market-maker inventories:

$$Cov(Ret_t, Ret_{t-k}) \quad = \quad -e^{-(k-1)\lambda\Delta t} \cdot \frac{\mu^2}{2\lambda} \cdot \beta_w^2 \cdot \left(\frac{1-e^{-\lambda\Delta t}}{r+\lambda}\right)^2.$$

The autocorrelation of returns is more complicated than the autocorrelation of market-maker inventories,

$$Corr(Ret_t, Ret_{t-k}) \quad = \quad -e^{-(k-1)\lambda\Delta t} \cdot \frac{\frac{\mu^2}{2\lambda} \cdot \beta_w^2 \cdot \left(\frac{1-e^{-\lambda\Delta t}}{r+\lambda}\right)^2}{\frac{\mu^2}{\lambda} \cdot \beta_w^2 \cdot \frac{1-e^{-\lambda\Delta t}}{(r+\lambda)^2} + \sigma_w^2 \Delta t}, \tag{27}$$

because of the variance due to the fundamental value component, $\sigma_w^2 \Delta t$. Given σ_w^2 is in the numerator of β_w^2, σ_w^2 is in all terms in the numerator and denominator and the autocorrelation of returns is independent of σ_w. In contrast, μ only appears in one of the denominator's terms in (27). Therefore, the autocorrelation of returns is increasing in μ. This enables the autocorrelation of returns to help identify μ, an identification not possible using the autocorrelation of market-maker inventories.

Overall, in the simplified version of the model with one class of slow investors, four parameters need to be identified: λ, β_M, μ, and β_w. The above discussion focuses on how the four equations for the variance and autocorrelations of market-maker inventories and returns [(24), (25), (26), and (27)] can be used to identify the model: the autocorrelation of market-maker

[25] By setting $\rho=0$ in this simplified version of the model, the variance due to new shocks does not include a component due to the shared effect.

inventories identifies λ; the variance of market-maker inventories helps identify β_M; and the variance and autocorrelation of returns help identify μ and β_w.

The simplified model only has one state variable, namely, the gap process. Both prices and market-maker inventories are proportional to the gap process (see (13) and (12)). Therefore, it is not surprising that observing market-maker inventories and returns are sufficient to identify the model without needing to use information about the slow investors.[26] This intuition holds for the general model.[27]

In addition to variances and autocorrelations of market-maker inventories and returns, the model provides the lead-lag cross-autocovariances of returns and market-maker inventories. However, the $Cov(Y_t, Y_{t-k})$ shows that

$$Cov(MMInv_t, MMInv_{t-k}) \cdot Cov(Ret_t, Ret_{t-k}) = Cov(MMInv_t, Ret_{t-k}) \cdot Cov(Ret_t, MMInv_{t-k}).$$

Therefore, once the autocorrelation of returns and market-maker inventories are known, the product of the lead-lag dynamics between returns and inventories is determined.[28] Hence, we focus on the covariance of returns with past market-maker inventories:

$$Cov(Ret_t, MMInv_{t-k}) \quad = \quad e^{-(k-1)\lambda\Delta t} \cdot \frac{\mu^2}{2\lambda} \cdot \beta_M \cdot \beta_w \cdot \frac{1-e^{-\lambda\Delta t}}{r+\lambda}.$$

Hence, dividing this autocovariance by the standard deviations of returns and market-maker inventories gives the cross-autocorrelation of returns with past market-maker inventories:

$$Corr(Ret_t, MMInv_{t-k}) \quad = \quad e^{-(k-1)\lambda\Delta t} \cdot \frac{\sqrt{\frac{\mu^2}{2\lambda}} \cdot \beta_w \cdot \frac{1-e^{-\lambda\Delta t}}{r+\lambda}}{\sqrt{\frac{\mu^2}{\lambda} \cdot \beta_w^2 \cdot \frac{\left(1-e^{-\lambda\Delta t}\right)}{(r+\lambda)^2} + \sigma_w^2 \Delta t}}. \quad (28)$$

While the structure of the above autocorrelation (returns and lagged market-maker inventories) share some similarity to the structure of the autocorrelation

[26] Note that while the model is identified using only market-maker inventories and returns, the parameters that are identified are less economically interesting. β_M, μ, and β_w are not exactly identified because μ is always multiplied by β_M or β_w. β_M and β_w share the same denominator (see (22)). Therefore, the four equations—(24), (25), (26), and (27)—identify λ, σ_w (the numerator of β_w), μ divided by the sum of the risk-aversion-adjusted masses of market makers and fast investors (the denominator of β_M and β_w), and the risk-aversion-adjusted mass of market makers (the numerator of β_M). Put another way, $\mu \cdot \beta_M$ and $\mu \cdot \beta_w$ are identified, but the values of μ and the betas are not separately identified. The source of μ, β_M, and β_w not being fully separable can be seen from (21), where the dynamics and market-maker inventories and returns based on the gap process are scaled by β_M and β_w. Note that in (21) the dynamics of the slow investors trading in row 8 is not scaled by the betas. Therefore, the variance of the slow investor trading can be used to separately identify μ, which then identifies β_M, and β_w (see also (11)). Because this section focuses on parsimoniously providing intuition for how the various moments help the identify the model, we do not write out moments based on slow investor trading.

[27] As long as a sufficient number of lagged autocorrelations are used relative to number of slow investor classes (which determines the dimension of the state variable), the dynamics of market-maker inventories and returns at different lags provide enough information to characterize the multidimensional state variable.

[28] This relation does not hold in the more general model with multiple classes of slow investors.

of returns (by themselves), it differs in that it is decreasing in σ_w. While the cross-autocorrelation of returns with past market-maker inventories is not needed for identifying the simplified model, the fit of this cross-autocorrelation can be thought of as an overidentification test in the empirical estimation.

3. Estimation and Results

3.1 Data

Our data start in January 1999 and end in December 2005 and come from four data sets:

- An internal New York Stock Exchange ("NYSE") database named the Specialist Summary File (or "SPETS") contains specialists' closing inventory positions for each stock at the end of each day. The NYSE assigns one specialist per stock and a given specialist is responsible for making a market in approximately ten stocks. See Hasbrouck and Sofianos (1993) for further discussion of the SPETS database.
- An internal NYSE database named the Consolidated Equity Audit Trail Data (or "CAUD") contains the number of shares bought and sold by retail (individual) investors, for each stock, over each day. In addition, the CAUD database provides trading volume. See Kaniel, Saar, and Titman (2008) for further discussion of the CAUD database.[29,30]
- The Trades and Quotes ("TAQ") database provides daily closing mid-quotes prices. Prices and returns in this paper are measured at the mid-quote to avoid bid-ask bounce. All prices are adjusted to account for stock splits and dividends.
- The Center for Research in Security Prices ("CRSP") provides the number of shares outstanding (used to calculate market capitalizations) and information necessary to adjust prices for stock splits/dividends.

Before discussing the details of the data, it is worthwhile to provide some context. During our sample period, 80% of trading occurred on the NYSE. Historically, the NYSE assigned one market maker (called a "specialist") to each stock. While the designation of a single market maker is relatively unique to the NYSE, the fundamental economic forces related to limited risk-bearing capacity for liquidity provision remain the same. It is likely that other investors, for example, hedge fund traders and, more recently, high-frequency traders, compete with the specialist by placing limit orders to supply liquidity.[31]

---

29 The investor classification in CAUD—together with market clearing—implies that the number of shares bought/sold by the market maker equals the sum of the number of shares bought/sold by the retail investors and institutions.

30 We convert market makers' inventory positions and retail investor net trades to U.S. dollars (both variables are originally in number of shares). For each stock, we multiply the number of shares by the stock's sample average price so as not to introduce price changes directly into the trading variables.

31 Hendershott and Moulton (2011) show the NYSE's market structure changes after our sample period (2006–2007) leading to a reduced role for the specialist and a decline in the NYSE's share of trading. This evolution highlights

Using the retail trading data from the NYSE has pros and cons similar to using the specialist data. The data represent a large, comprehensive sample of trades. However, there exist retail trades with broker dealers who internalize orders and trades on markets other than the NYSE.[32] As discussed above, the data on slow investors, which include retail traders, are not required to identify the model's price dynamics but are useful for identifying more economically intuitive parameters in the model.

We start with the 2,357 common stocks that can be matched across the NYSE, TAQ, and CRSP databases. We construct a quasi-balanced panel of data to ensure the results are comparable throughout time. To do this, a stock's data need to be available at the beginning and end of our sample period. Stocks with an average share price of less than \$5 or larger than \$1,000 are removed from the sample. The final sample consists of 689 actively traded stocks.[33]

Idiosyncratic variables. We focus on idiosyncratic components of our variables for several reasons. First, the autocorrelation of stock market returns is not statistically significantly different from zero. Therefore, unlike for individual stocks, there is little evidence to suggest marketwide pricing errors. This could arise from the risks associated with marketwide return shocks being able to be hedged with highly liquid index products. Second, unlike the market-maker inventory dynamics for individual stocks, the systematic marketwide component of the market-maker inventories exhibits little autocorrelation.

For each return and trading variable, we construct a common factor equal to the market capitalization weighted average of the underlying variable. We regress each variable on its common factor and save the residual as the corresponding idiosyncratic variable. For notational simplicity, we omit any subscripts or superscripts referring to "idiosyncratic" and, for example, use $MMInv_t$ to denote the idiosyncratic component of market-maker inventories. After removing the marketwide components, the contemporaneous pairwise correlations (across firms) are 0.034, 0.009, and 0.036 in market maker inventories, retail flows, and returns, respectively.

This idiosyncratization procedure has a strong effect on returns (not surprisingly) but has only a very weak effect on the trading variables. For example, in the cross-section, the variance of the idiosyncratic components

a potential weakness of our data, as well as some strengths. On the positive side, the NYSE specialist system that we study is the market structure underlying much of the data used in modern asset pricing. Comprehensive data on the trades and positions of other liquidity suppliers who compete with (or replace) the specialists are not available. When or whether such data may become available is unclear.

32 Our retail trading and market-maker inventory data may not be comprehensive for all such market participants; for example, other investors provide liquidity. If not, as long as our data are representative of market participants, then all of our estimation results remain unchanged, except for (a) our estimate of the market-makers' share of the immediate risk aversion-adjusted mass (β_M) is likely to be a lower bound and (b) the risk masses of the retail-investor classes (μ_r's), as fractions of the risk mass of all slow investors, become lower bounds.

33 The sample is similar to the one used in Hendershott and Menkveld (2014). For a more detailed characterization of the stocks, please see that paper.

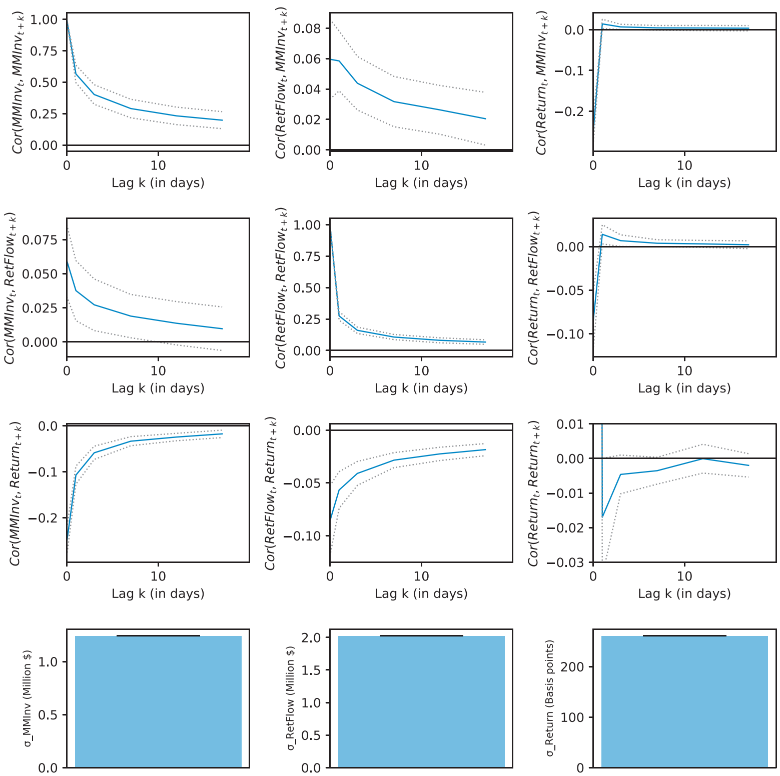

Figure 1
Empirical moments
The upper nine plots show the empirical autocorrelations and cross-autocorrelations with 95% confidence bands. Standard errors are based on a block bootstrap methodology as discussed in footnote 37 and in Internet Appendix B. The lower three plots show empirical standard deviations.

are 97.4% of total variance for $MMInv_t$ and 99.9% for $RetFlow_t$. *Not* idiosyncratizing trade variables likely affects model estimates only mildly, yet we prefer to use the idiosyncratic versions to not introduce bias. The model focuses on nonsystematic effects; order flows and positions due to (marketwide) systematic effects are removed by the procedure described above.

Figure 1 plots our three variables' autocorrelations and the lead-lag correlations among the variables up to a lag of 20 days. For ease of exposition, we refer to the lead-lag correlations as cross-autocorrelations. The upper nine plots show the (cross) autocorrelations of all possible ordered pairs of the three series. Note that contemporaneous correlations are shown at lag zero on the six off-diagonal plots. The lower three plots show the standard deviation of each series.[34] These 12 plots in Figure 1 illustrate the multivariate autocovariance function for all series with lags ranging from 0 days to 20 days (monthly). The plots summarize the dynamics of our data series. The plots show the cross-sectional average moments along with their 95% confidence bands based on a block bootstrap methodology as discussed in footnote 37 and in Internet Appendix B. As a robustness check, Internet Appendix G shows parameter moments based on bootstrap means.

The figure has some notable and statistically significant patterns. First, the standard deviation of market-maker inventories is \$1.25 million (see plot (4,1)). Inventories initially decay rather quickly as the first-day autocorrelation is 0.58 (see plot (1,1)). After a day, they decay slowly and end with a 20-day autocorrelation of 0.2. Second, the standard deviation in retail flows is \$2.0 million (see plot (4,2)). Similar to market-maker inventories, they decay extremely quickly on the first day (even more quickly than do market-maker inventories) and then slowly over the following 19 days (see plot (2,2)). Third, the standard deviation of idiosyncratic daily returns is 2.5% (or 250 basis points [bps]) (see plot (4,3)). The return autocorrelations are negative throughout the majority of the first 20 days suggesting that at least part of the original pricing error is persistent (see plot (3,3)). Focusing again on plot (3,3), the average return autocorrelation is -0.0056 with a standard deviation of 0.0019 across our block bootstrap draws (indicating statistical significance at all conventional levels).

The figure further reveals strong cross-autocorrelations. First, market-maker inventories and retail flows are positively correlated, both contemporaneously and through time (see plot (2,1) and (1,2)). This pattern is consistent with our model. Periods when market makers are long securities correspond with periods when retail investors are buying. Such a pattern is consistent with market makers holding securities for slow retail investors to later purchase.

Second, there is a strong negative correlation between the market-maker inventories and contemporaneous returns (-0.25) that turns to a modest positive correlation with future returns (but is basically zero after day 10). This pattern suggests market makers are compensated for intermediation (see plot (1,3)). They purchase securities cheaply to sell at higher future prices. Such selling is consistent with the current return correlating steadily less negatively with future inventories (see plot (3,1)).

Third, a negative current return correlates with retail investors buying contemporaneously as well as with continued retail buying in days to come

34 To equally weight across stocks, we standardize all empirical variables (market maker inventories, retail flows, and returns) to have equal variance across stocks by dividing each variable (for each stock) by that variable's variance within that stock. All variables are then rescaled by the average of that same variance across all stocks to ensure the variance of the full sample equals the average variance across stocks.

(see plot (3,2)).[35] This is consistent with a positive target shock that makes the more-attentive retail investors buy now, while slower retail investors buy at a later time. While inattention in our model causes retail investors to have lower utility, retail investors benefit in monetary terms because as a group they appear to buy below fundamental values. Even those who arrive late seem to buy at depressed prices. To see this point, cumulate the strong contemporaneous negative return with modest future positive returns shown in plot (2,3).

Figure 1 also illustrates the similarities between market-maker inventories and retail flows. The correlation of these variables with returns is similar (compare plot (1,3) with (2,3) and also compare plot (3,1) with (3,2)). In addition, market makers inventories and retail flows are positively correlated, see plots (1,2) and (2,1). In our model, these correlations follow from market makers and retail traders both trading against price pressure, that is, selling when the pricing error is positive. Trading against price pressure is often an important component when defining liquidity provision. However, liquidity provision typically also involves temporarily holding a suboptimal position and profiting from the pricing error (i.e., when the pricing error is positive, own less of an asset than when the pricing error is zero or negative.) In our model, market makers hold suboptimal positions in order to profit from the pricing error, but retail traders (who are present) do not. Therefore, our model illustrates how the empirical correlations between retail flows and returns—consistent with the liquidity provision shown in Kaniel, Saar, and Titman (2008)—can arise from different motivations than those of market makers.[36]

3.2 Results

A standard MLE procedure is used to estimate the model's deep parameters. For a particular value of the parameters, the likelihood is evaluated recursively (through time) using the Kalman filter (Durbin and Koopman, 2012, Ch. 7). This likelihood is then optimized with respect to these parameters by using a standard steepest-ascent method. The method requires picking starting values, which is done by matching a subset of autocovariances in the data. Please see Internet Appendix A for a detailed discussion.

We present "pooled" estimates for all stocks in our sample, as well as for sized-based subsamples of stocks. Stocks are divided into large, medium, and small terciles. The large tercile is divided in half (upper-half and lower-half). When performing a given estimation, firm data are stacked. A sufficient number of empty observations are inserted between the firms so as not to affect the

35 In our model, if $\rho = 0$, then hedging shocks are uncorrelated with fundamental-value innovations implying a zero contemporaneous correlation of retail flows and returns, which is inconsistent with the data. This implies that the contemporaneous correlation of retail flows and returns plays an important role in identifying ρ.

36 In our model, inattentive traders (such as retail traders) who are present in the market, trade directly to their target portfolios. If this modeling assumption is relaxed, as in Duffie (2010), then retail traders engage in some liquidity provision as well as trading due to hedging needs and inattention.

Table 2
Parameter estimates

	All	*Large stocks*		*Medium*	*Small*
	stocks	*Upper-half*	*Lower-half*	*stocks*	*stocks*
A. Risk masses of slow institutional investors					
μ_{di}	151 *** (13.9)	589 *** (98.4)	142 *** (1.13)	62.5 *** (3.01)	12.8 *** (1.23)
μ_{mi}	25.1 *** (2.32)	78.4 *** (13.1)	17.5 *** (1.44)	9.29 *** (0.51)	3.85 *** (0.45)
μ_{qi}	7.76 *** (0.69)	40.8 *** (6.65)	8.41 *** (0.18)	2.18 *** (0.16)	0.65 (0.50)
B. Risk masses of (slow) retail investors					
μ_{dr}	1.63 *** (0.08)	3.92 *** (0.17)	1.18 *** (0.049)	0.494*** (0.03)	0.182*** (0.04)
μ_{mr}	4.97 *** (0.37)	12.8 *** (1.22)	3.08 *** (0.28)	1.39 *** (0.10)	0.62 *** (0.11)
μ_{qr}	2.03 *** (0.19)	6.63 *** (1.25)	2.74 *** (0.38)	0.916*** (0.07)	0.036 (0.06)
C. Deep parameters					
β_M	0.0082 *** (0.0007)	0.0042 *** (0.0007)	0.0077 *** (0.0003)	0.0089 *** (0.0005)	0.0209 *** (0.0019)
β_w	0.0933** (0.037)	0.0522*** (0.007)	0.0371 (0.050)	0.2570** (0.117)	0.5370 (0.409)
D. Volatility related to returns, market-maker inventories, and retail flows					
σ_w	222 *** (17.4)	125 *** (11.1)	232 *** (18.9)	223 *** (16.7)	255 *** (19.8)
σ_{e_M}	0.385 *** (0.080)	1.42 *** (0.133)	0.538 *** (0.042)	0.174 ** (0.040)	0.001 (0.001)
σ_{er}	1.58 *** (0.034)	3.43 *** (0.074)	1.07 *** (0.021)	0.50 *** (0.01)	0.22 *** (0.01)
E. Shared component					
ρ	−0.223*** (0.022)	−0.233*** (0.015)	−0.280*** (0.020)	−0.222*** (0.024)	−0.234*** (0.039)
# of stocks	689	115	115	229	230
# of obs.	1,206,935	201,984	201,987	402,169	400,795

This table presents the maximum likelihood parameter estimates and their standard errors. We consider "All stocks" as well as four size-based subsamples labeled "Large-upper," "Large-lower, "Medium," and "Small" (the large tercile has been divided in half.) Subscripts: "d" daily; "m" monthly; "q" quarterly; "'slow institutional investors; and "r" slow retail investors. Idiosyncratic noise in dividends (σ_w); market-maker inventories (σ_{e_M}); and retail flows (σ_{e_r}). Standard errors are shown in parentheses and are based on a block bootstrap methodology as discussed in footnote 37 and in Internet Appendix B. *p <.1; **p <.05; ***p <.01.

lead-lag relationships inherent within a given firm's data. Also, because market makers and retail traders are likely to trade for reasons other than those in our model, our estimation allows for shocks to the market-maker inventories and retail investor trading that are independent of the model's shocks. The standard deviations of these shocks are σ_{e_M} and σ_{e_r}, respectively.

Table 2 presents the estimation results for our full sample ("All Stocks") as well the size-based subsamples. All but five parameter estimates are significant at the 5% level.[37] We now discuss important results related to Table 2.

First, focusing on "All Stocks," the slow institutions' risk masses are, perhaps not surprisingly, a lot higher than those of the retail investors. The slow

[37] Throughout this paper, standard errors are computed using a block bootstrap procedure. Each firm is mapped into 1 of 30 industries based on Kenneth French's website. We then draw one firm from each industry, create a sample based on this subset of firms, estimate the model, and save the parameter and figure values. The procedure is repeated 100 times. Internet Appendix B gives an additional discussion.

institutions' risk masses are 151, 25.1, and 7.76 for daily, monthly, and quarterly frequencies, while the slow retail investors' risk masses are 1.63, 4.97, and 2.03 for the same frequencies.[38] Note further that comparing within investor type, there is relatively more retail risk mass at the quarterly frequency than at the daily frequency. This confirms our intuition that institutions, even the ones who are slow, are still relatively faster than retail investors. However, the results do show that some attentive retail investors visit the market once a day on average and some relatively inattentive institutions who visit the market once a quarter on average.[39]

Second, our estimate of β_M is 0.0082 for "All Stocks." If $\gamma_M = \gamma_F$, then market makers make up 0.82% of the "fast" risk mass, while fast institutions make up the other 99.18%. If the two risk aversion parameters are similar, but not equal, we can conclude that market makers are small relative to fast institutions. Cross-sectionally, we see the role of market makers (β_M) increases monotonically from large-upper stocks (0.0042) to small stocks (0.0209), consistent with Hendershott and Menkveld (2014).

The idiosyncratic component of dividend risk (σ_w) generally declines with firm size. For small stocks σ_w=2.55% and for large-upper stocks σ_w=1.25%. Similarly, the sensitivity of the pricing error to the gap process, β_w, generally declines with firm size. This is consistent with trading costs being lower in larger stocks and with results in Hendershott and Menkveld (2014).

Third, the risk-bearing capacity of fast participants (market makers plus fast institutions) is substantial. A one-standard-deviation shock to each slow investor class corresponds to a shock to the fast investors target portfolios of \$192 million (the sum of the μ's).[40] Weighting these shocks by the reciprocal of their corresponding λ and multiplying the sum by β_w yields a pricing error of 1.3%, that is, substitute (22) into (12) and set r=0. A one-standard-deviation shock for large stocks is almost \$500 million, with a resultant price pressure of 1.7%. For small stocks, a one-standard-deviation shock is only \$18 million,

[38] The risk masses also correspond to the standard deviations of the hedging shocks in millions of dollars. Cross-sectionally, the shocks are declining in firm size.

[39] The retail risk-mass estimates imply approximately 20% of retail is at the daily frequency. When interpreting this fraction, a number of factors must be considered. First, loosely speaking, the total mass of slow traders is identified by the market makers inventories. The retail flow data identify how much of the slow trader mass is retail, while the residual represents institutional (nonretail) slow traders. The relative sizes of institutional and retail slow traders can be seen by comparing the $\mu_{.,i}$'s to $\mu_{.,r}$'s. At a daily frequency, the slow institutions are roughly 100 times larger than retail investors, except for the small stocks, where the ratio is about 50 times. Hence, the retail traders are small at a daily frequency; though, at a quarterly frequency, the slow institutions are only 2–18 times larger than retail investors. Second, the less sophisticated retail traders' orders have historically been paid for (payment for order flow) and not sent to the NYSE (see a discussion in Kaniel, Saar, and Titman (2008)). Third, retail trading includes balanced buying and selling (which nets to zero in our retail flow data). In addition, there is retail flow outside of our model that we estimate with the σ_{e_r} parameter. Therefore, while it is true that roughly 20% of the total mass of retail investors arrive daily, our model and data only identify a subset of all retail trading.

[40] Note that this argument relies on shocking each investor class by its standard deviation. If instead, one is interested in a one-standard-deviation shock to the sum of target portfolios, then there is some diversification to be accounted for. The relative differences across small, medium, and large stocks remain mostly unaffected as the shared components are approximately the same size across the size-based subsamples (see the ρ's in Table 2).

with a resultant price pressure of 0.8%. Comparing price pressures across size-based subsamples suggests that the price impact of a $1 million shock to fast investors corresponds to 0.35 bps for large stocks and 4.4 bps for small stocks.[41]

Fourth, we can measure the correlations of shocks across firms (intuitively, this correlation is closely related to the correlation of our returns discussed earlier). Adding an additional parameter for the shock covariance across firms to the estimation is difficult given the number of parameters already being estimated and the smaller sample size in the bootstrap estimation (30 firms at a time for the standard errors). However, we can calculate the shock correlation across firms implied by our estimates. For the estimation with “All Stocks,” the shock correlation is 0.0137, though not reported directly in the table.

Finally, note that shocks to the target holdings of slow investors correlate negatively with fundamental value changes ($\rho = -0.223$ for “All Stocks”). This negative correlation amplifies the security’s volatility as, for example, a sudden drop in the security’s fundamental value coincides with contemporaneously higher target levels for slow investors and, therefore, with lower target holdings for fast investors. In other words, fast investors want to sell their securities as fundamental values are dropping. On average, the fast investors cause a negative pricing error, which adds to the price drop. Because this shock applies equally to slow investors at all frequencies, we refer to the negative correlation as a “shared component.” As discussed in Section 1.1, the shared component could arise from fast investors having larger shocks than slow investors. Such an imbalance in shock size would then lead to the permanent component in prices being positively correlated with the pricing error.

Figure 2 illustrates the model’s fit using “All Stocks.” We plot the empirical autocorrelations, cross-autocorrelations, and standard deviations (same as those shown in Figure 1) as well as the model-implied counterparts. This figure gives a visual overview of the estimation results shown in Table 2’s first column.[42] Additional Estimation results for the size-based terciles are given in Internet Appendix E.

3.3 Characterization of pricing errors

The estimated model characterizes the pricing errors, their effect on returns, and how slow institutions and retail investors contribute to them. Overall, we find pricing errors are significant and long lasting.

[41] The pricing error for a one-standard-deviation shock being larger for large stocks compared to small stocks differs from Hendershott and Menkveld (2014). As we will see in Section 3.3, most of the pricing error in this paper is due to very persistent shocks whereas in Hendershott and Menkveld (2014) the shocks are not persistent. This difference could cause the statistical estimation in Hendershott and Menkveld (2014) to incorrectly identify long-lived pricing errors as permanent price changes.

[42] The empirical returns are computed as the first difference of log prices. We therefore implicitly assume that log-price differences are normally distributed when fitting the model. In particular, *log*-prices can become negative as is the case in the model. Note further that the model’s dividend shocks correspond to log fundamental-value changes in the data.

The upper panel of Figure 3 decomposes the steady-state pricing error variance into four components based on (E.10) from Appendix E. The largest component is due to quarterly slow investors and amounts to a standard deviation of 2.547%. Daily slow investors only contribute 0.097% to the overall standard deviation, while monthly slow investors contribute 1.574%. The difference in contributions stands in stark contrast to the risk masses of the three classes; the risk mass of daily investors (μ_d) is much larger than the risk mass of quarterly investors (μ_q), as shown in Table 2. The reason for the wedge between risk masses and contributions is that quarterly investors' hedge shocks last 60 days, whereas the daily investors' shocks last only a single day. Figure 3 further shows that the shared component in pricing errors is sizable with a 1.105% standard deviation.

The lower panel of Figure 3 illustrates the decay in pricing errors by plotting their autocorrelation function. The plot is based on (E.11) from Appendix E. The lower panel clearly shows that pricing errors are very persistent and decay only slowly over time. After a month (i.e., trading 20 days) almost two-thirds remain. The reason is that the pricing errors are dominated by the quarterly slow investors, as the decomposition in the upper panel clearly shows. The half-life of a pricing error shock is just over 31 days or 6.2 weeks and also can be seen in the lower panel.

The persistence of pricing errors cause them to substantially affect daily, monthly, *and* quarterly returns. Figure 4 illustrates this observation by decomposing returns into three components based on (E.12) from Appendix E. Fundamental-value innovations constitute the largest component of returns at all frequencies. Its standard deviation is 2.224%, 10.193%, and 17.655% for daily, monthly, and quarterly returns, respectively. The standard deviations of the other components range from 0.11% to 0.707% for daily returns, from 1.317% to 2.462% for monthly returns, and from 2.351% to 3.034% for quarterly returns.

We use Figure 4 to calculate the relative contribution of pricing errors to idiosyncratic return variance. For daily returns, we see $(11.0^2+70.7^2)/222.4^2 = 9.4\%$ indicating that pricing errors account for 9.4% of daily idiosyncratic return variance. Similar calculations show that pricing errors account for 7.0% and 4.5% of respective monthly and quarterly variances.

Note that the relative sizes of the *legacy error reduction* components represent the most salient differences between the daily and quarterly returns. The strong error persistence reduces the amount of pricing error eliminated over all time periods, with the impact being higher the shorter the time period. This makes the legacy error reduction a small component in daily returns and a modest component in quarterly returns. This feature is also the root cause for why first-order autocorrelations are more negative for monthly returns than for daily returns in Table 1 (see also the discussion in Appendix A).

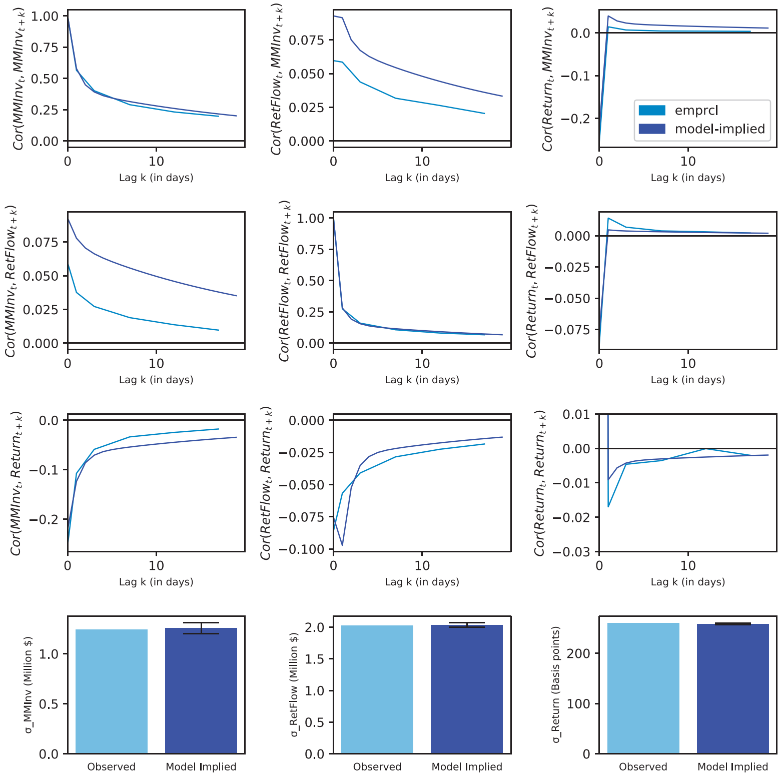

Figure 2
Empirical and model-implied moments
This figure illustrates the model's fit. The light-blue lines in the top nine plots and light-blue shaded bars in the bottom three plots are the empirical moments. The dark-blue lines and dark-blue bars are the model-implied moments. Parameters are estimated with maximum likelihood. This model features slow investors who arrive at daily, monthly, and quarterly frequencies (on average). The standard errors for the dark-blue lines in the top nine plots are shown in Internet Appendix H. The standard errors for the model-implied values in the bottom-three plots are shown in this figure. Standard errors are based on a block bootstrap methodology as discussed in footnote 37 and in Internet Appendix B.

3.4 Counterfactual analysis

Our structural model allows us to conduct some counterfactual analyses. We consider two dimensions. First, we assume the risk aversion of all fast investors is either $\frac{1}{2}$ or 2 times the value implied by the structural estimation. Equation (22) shows the role risk aversion plays in β_w and the estimated values for this parameter are shown in Table 2. Half or double the risk aversion is implemented by adjusting β_w by halving or doubling it. These counterfactual scenarios could arise if changes to regulations (in an attempt to influence financial stability by reducing speculation by banks and institutions) affect the risk aversion of fast investors in our model.

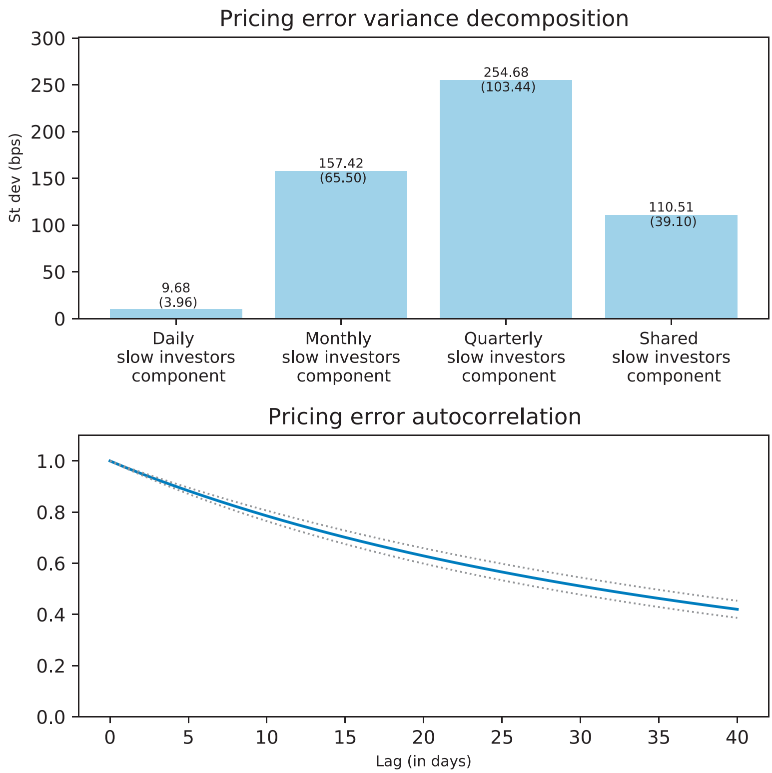

Figure 3
Pricing error magnitude and duration
The top panel of the figure illustrates the magnitude of the pricing errors along with a decomposition across the frequencies at which slow investors arrive (i.e., daily, monthly, and quarterly). The bottom panel of the figure shows how pricing errors decay over time. These graphs are based on parameter estimates from Table 2, Column "All." The upper panel shows standard errors in parentheses, and the lower panel shows 95% confidence bands. Standard errors are based on a block bootstrap methodology as discussed in footnote 37 and in Internet Appendix B.

Second, we change the arrival intensities of the *institutional* slow investors in one of two different ways. In case A, we assume all slow institutional investors arrive once a day (on average): $\mu'_{di} = \mu_{di} + \mu_{mi} + \mu_{qi}$ with $\mu'_{mi} = 0$ and $\mu'_{qi} = 0$. In case B, we assume the daily slow institutions become fast institutions: $\mu'_{di} = 0$, while the other slow institutions remain unchanged. These changes to the institutions' slowness could arise from investments in technology, enabling more frequent attention. The two cases represent the slowest institutions becoming faster and the least-slow slow institutions becoming fast, respectively.

To quantify the effects of changing risk aversion or risk masses, we record the fraction of pricing errors in daily, monthly, and quarterly idiosyncratic returns

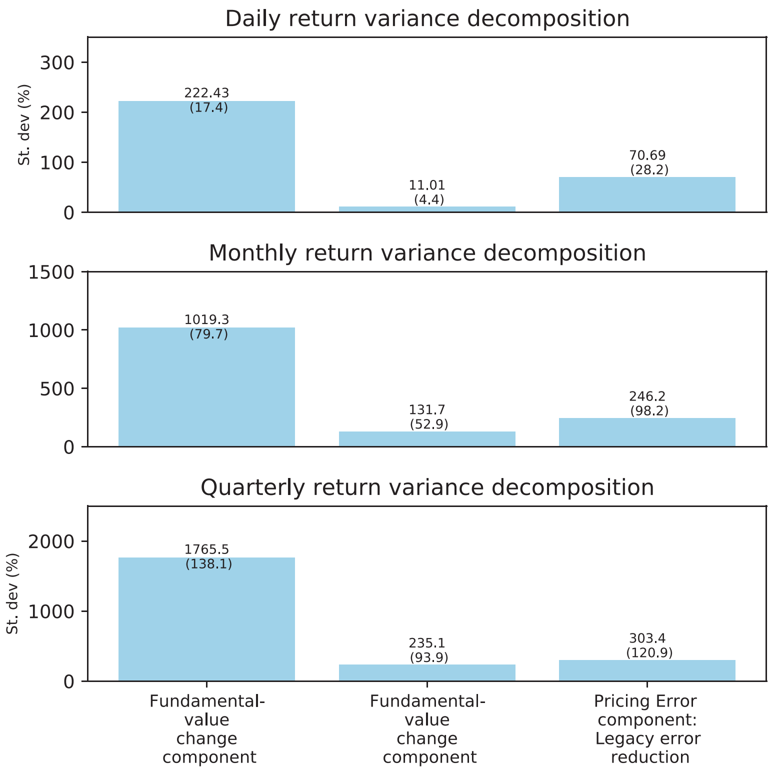

Figure 4
Return volatility decomposition
The graphs illustrate how total idiosyncratic return volatility can be decomposed into a fundamental-value change component and two pricing error components (a legacy-error reduction component and a new target portfolio shocks' component). The upper panel illustrates the decomposition for daily return volatility; the middle panel illustrates monthly return volatility; and the lower panel illustrates quarterly return variance. These graphs are based on estimated parameters using " All Stocks." Standard errors are shown in parentheses and are based on a block bootstrap methodology as discussed in footnote 37 and in Internet Appendix B.

(the values calculated using results shown in Figure 4). Figure 5 summarizes the results of the counterfactual analysis.

Starting with the top panel in Figure 5, we see that pricing errors account for 9.4% of return variance ("base case" as shown in column 1). If the risk aversion of the fast investors falls in half (these investors become more risk tolerant), we see pricing errors account for only 2.5% of return variance (column 2). If these investors' risk aversion doubles, pricing errors account for 29.3% of return variance (column 3). The results are not surprising. As fast investors become more risk averse, they require more compensation (larger pricing errors) to trade.

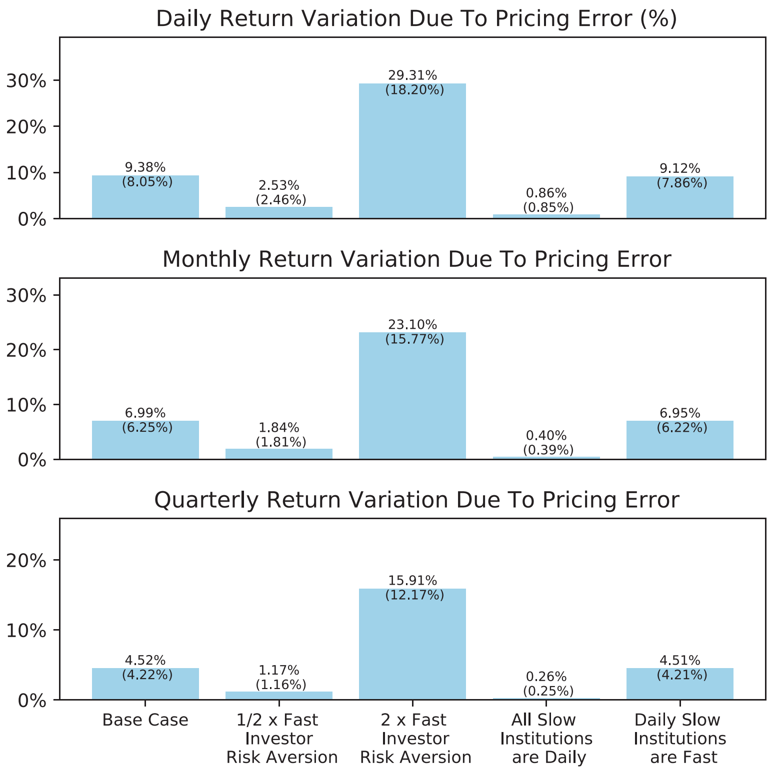

Figure 5
Counterfactual analysis
We report the fraction of return variance due to pricing errors for daily returns (top panel), monthly returns (middle panel), and quarterly returns (lower panel). The counterfactual analysis considers the fast investors' risk aversion to either fall by 50% or double (columns 2 and 3). Also, we consider a scenario in which all slow institutions arrive daily on average (column 4) and a scenario in which daily slow institutions become fast (column 5). Standard errors are shown in parentheses and are based on a block bootstrap methodology as discussed in footnote 37 and in Internet Appendix B.

Next, we consider all slow institutions arriving daily (on average). The net result is zero risk mass at monthly and quarterly frequencies ($\mu^{'}_{mi}=0$ and $\mu^{'}_{qi}=0$). As can be seen in column 4, the fraction of return variance due to pricing errors goes to almost zero (it is 0.9% in the top panel). However, if the daily slow institutions (only) become fast investors, the reduction in pricing errors is negligible as column 5 is 9.1% versus the column 1 value of 9.4%. We see similar patterns of results when looking at the monthly and quarterly return variances in the second and third panels in Figure 5.

The counterfactual analysis illustrates the relevance of accounting for arrival intensities of investors when explaining price pressure. We see that the perdollar price pressure scales with the *inverse* of intensity (i.e., tending to infinity as the intensity tends to zero). For example, in our base case, the risk mass of daily slow investors is 16 times larger than the risk mass of quarterly slow investors, yet the price pressure they command in four times lower.[43]

4. Conclusion

We analytically solve a structural model with inattention and estimate its parameters using the dynamics of NYSE market-maker inventories, retail order flows, and prices. The model and trade data enable identification and measurement of pricing errors' role in stock return volatility. We find that pricing errors account for 9.4%, 7.0%, and 4.5% of the respective daily, monthly, and quarterly idiosyncratic return variances.

Our model and empirical approach can be applied to other data from a range of investor groups and over different time horizons. For example, even lower-frequency dynamics could be estimated using data from very long-term investors. Such data could be obtained from public SEC filings (13F) or private data providers, such as Ancerno. Our continuous-time model also can be translated into frequencies as high as a millisecond. Data from exchanges identifying high-frequency traders therefore also could be incorporated into our approach to examine these traders' roles in correcting or possibly causing pricing errors. An important component of extending our approach to other samples is to identify and measure market-maker inventories.

In the future, it also may be possible to add informational frictions to our model. These frictions could potentially help quantify the role attention plays in prices slowly adjusting to new information. Data sources on public news also could be incorporated to measure how attention varies with both market conditions and the arrival of information. At the lowest frequencies, macroeconomic variables could be added to study how they affect the duration of pricing errors. Finally, a more comprehensive understanding of stock return patterns could combine overreaction from the inattention and risk sharing (as in this paper) with potential underreaction in stock returns from endogenous information acquisition (see Sims (2003); Nieuwerburgh and Veldkamp (2009)).[44] These extensions provide examples of potentially important future work.

---

[43] The corresponding calculations are from the "All Stocks" column in Table 2: (151 + 1.63) / (7.76 + 2.03) = 16. Next, we see $16\times\left(\frac{1}{63}\right)=0.25$. Here, a calendar quarter consists of 63 days.

[44] Empirically and theoretically trades on average have a permanent price impact. This entirely comes from trading by informed traders causing permanent price changes. If these traders acquire information over time, prices may adjust slowly. Trading by those without information, such as studied in our paper, has no permanent price impact. Rather, it only has a transitory price impact.

Appendices for

"Asset Price Dynamics with Limited Attention"

A. Pricing Error Persistence and Return Autocorrelation

This section explores how pricing errors relate to return autocorrelations. If the first-order autocorrelation of returns is highly negative, then pricing errors must be large relative to fundamental value changes. However, the reverse need not be true. If pricing errors are persistent then they can be relatively large *while* short horizon return autocorrelations can be small. This may explain why pricing errors have largely been overlooked in the literature. Daily return autocorrelations are typically small and one might (erroneously) conclude that pricing errors can safely be ignored. Our paper shows that such errors are economically large for actively traded U.S. equities.

To examine pricing errors, assume that daily (log) prices, say mid-quotes, consist of two unobserved components: a martingale m_t plus an error term s_t. The first-order autocovariance of *daily returns* is

$$\mathrm{cov}(w_t+s_t-s_{t-1}, w_{t-1}+s_{t-1}-s_{t-2}) = -\left(1+\rho_{s,2}-2\rho_{s,1}\right)\sigma_s^2 < 0, \tag{A.1}$$

where w_t is the martingale innovation and $\rho_{s,i}$ is the ith order autocorrelation in the pricing error s_t. Assuming that w_t and s_t are uncorrelated yields the following expression for return variance:

$$\mathrm{var}(w_t+s_t-s_{t-1}) = \sigma_w^2+2\left(1-\rho_{s,1}\right)\sigma_s^2. \tag{A.2}$$

The first-order autocorrelation of daily returns is therefore

$$\rho_{r,1} = -\frac{\left(1+\rho_{s,2}-2\rho_{s,1}\right)\sigma_s^2}{\sigma_w^2+2\left(1-\rho_{s,1}\right)\sigma_s^2}. \tag{A.3}$$

Case A: Pricing errors are uncorrelated across days. If pricing errors are not persistent ($\rho_{s,1} = \rho_{s,2} = 0$), then the first-order autocorrelation in (A.3) becomes

$$\rho_{r,1} = -\frac{\sigma_s^2}{\sigma_w^2+2\sigma_s^2}. \tag{A.4}$$

When pricing errors are large relative to fundamental-value innovations, the first-order return autocorrelation is large and negative. When pricing errors are small relative to fundamental-value innovations (σ_s^2 is small relative to σ_w^2), the above expression is small, negative, and approximately equal to minus the ratio of σ_s^2 to σ_w^2. Finally, note that the pricing errors' relative size diminishes as one downsamples the data from a daily to, say, monthly frequency. To understand the effects of daily-to-monthly downsampling, notice that when σ_s^2 is small relative to σ_w^2, the denominator in (A.4) increases by a factor of 20 (approximately) while the numerator remains unchanged. This implies that downsampling makes the first-order return autocorrelations less negative the lower the sampling frequency.

The first-order autocorrelations of returns in Table 1 shows that the autocorrelation is more negative at a monthly frequency than at a daily frequency. The above downsampling logic demonstrates that empirical return autocorrelations are inconsistent with pricing errors being uncorrelated across days.

Case B: Pricing errors *are correlated* across days. Persistent pricing errors are difficult to detect in first-order return autocorrelations. This is perhaps best seen by considering the following limit:[45]

$$\lim_{\rho_{s,2}\uparrow 1} \rho_{r,1} = 0. \tag{A.5}$$

This limit shows that as pricing errors become persistent enough the first-order return autocorrelation approaches zero. Essentially, the pricing error begins to resemble a martingale so returns become uncorrelated.

How does pricing-error persistence affect return autocorrelations at different sampling frequencies? Downsampling mechanically reduces pricing error persistence, which can help disentangle longer-lived pricing errors from the martingale component of prices. For example,

45 Formally showing this limit requires additional assumptions about pricing error process; for example, its variance must remain finite.

shocks to pricing errors might live for several days causing high persistence at a daily frequency. If the shocks largely die out at longer (monthly) frequencies, the pricing errors sampled at such frequencies exhibit only moderate persistence. Downsampling therefore guarantees that at some point first-order return autocorrelations becomes substantially negative again. Therefore, pricing-error persistence can make these autocorrelations more negative at lower frequencies: the derivative of $\rho_{r,1}$ with respect to the sampling horizon can be negative.

We illustrate how this line of reasoning can generate the empirical patterns shown in Table 1. Let (daily) pricing errors decay exponentially with intensity 1/20 so that the expected duration is a month. Let both their (unconditional) standard deviation and the daily martingale innovation be 1%. Then simply applying (A.3) yields a first-order autocorrelation in daily returns of -0.01. This autocorrelation is however -0.05 when computed for monthly returns. These autocorrelations (-0.01 and -0.05) are quite close to those for the U.S. stock market data in Table 1. In our model, the slowly decaying pricing errors are generated by some inattentive investors who only participate in the market once a month on average.

B. Notation Summary

This appendix summarizes the notation used throughout our paper. We first describe the model's variables and parameters, next the data series, and finally the estimated parameters.

Variable	Description
Δt	The period length implied by the sampling frequency
ε_t	Vector with all the model's shocks in the period from $t-\Delta t$ to t
G_t	The gap between target and actual portfolio aggregated across all investors
λ_i	Poisson arrival intensity for investor class $i \in \{d,m,q\}$
	Note on subscripts: "d" daily; "m" monthly; "q" quarterly;
Λ_k	Diagonal matrix with the intensities of investor class $k \in \{i,r\}$ on the diagonal
m	Mass of investors
r	Risk-free rate
σ	Per-investor shock in target portfolio
Y_t	State vector that stacks all the model's unobserved and observed variables

Data	Description
$MMInv_t$	Market makers' inventory at time t
$RetFlow_t$	Retail investor order flow in the period from $t-\Delta t$ to t
$Return_t$	Asset's idiosyncratic (mid-quote) return in the period from $t-\Delta t$ to t

The 12 estimated parameters	Description
β_w	Asset's fundamental-value risk relative to total risk-absorption capacity (i.e., market makers' plus fast investors)
β_M	Market makers' risk-absorption capacity relative to total capacity (i.e., market makers' plus fast investors)
$\mu_{j,l}$	The size of the total target portfolio shock aggregated across all slow investors of class $j \in \{i,r\}$ and frequency $l \in \{d,m,q\}$. Note 1: There are six $\mu_{j,l}$ variables Note 2: $\mu := m\sigma$ is the mass of investors times average per-investor shock size Note 3: "i" institutional; "r" retail Note 4: "d" daily; "m" monthly; "q" quarterly;
ρ	Correlation of all slow investors target portfolio shocks and the asset's fundamental-value change
σ_w	Asset's fundamental risk (i.e., standard deviation of the asset's fundamental-value changes)
σ_{e_M}	Daily noise in market-maker inventories
σ_{e_r}	Daily noise in retail order flow

C. Imbalanced Shocks: Their Correlation with Balanced Shocks and Returns

This appendix illustrates permanent price dynamics[46] that can result from nonzero sum (or imbalanced) shocks to investors' target portfolios. Crucially, below we assume that permanent price changes are linear in imbalanced shocks. Permanent price dynamics may arise even if all investors are (always) attentive and present to trade. Therefore, for ease of illustration, assume only two classes of investors and that both classes are fast and attentive. In this setting, there is no need for market makers. For consistency with the model in the body of the paper, we use the same labels for the different classes of traders:

- Class 1 of traders (indexed n)
- Class 2 of traders (indexed F)

There are two (possibly correlated) Brownian motions:

- Z_t represents "balanced" shocks as in (equation reference in main text).
- I_t represents "imbalanced" shocks (new process in this appendix) .

The imbalanced shock is distributed across the investors' target portfolios by the function $f(I_t)$:

$$\begin{aligned} T_{n,t} &= \sigma_n Z_t + (1 - f(I_t)) I_t \\ T_{F,t} &= -\frac{m_n}{m_F}\sigma_n Z_t + \frac{m_n}{m_F} f(I_t) I_t \end{aligned}$$

Note: If $I_t = 0 \ \forall\ t$ (no permanent imbalance) then the sum of the target portfolios is zero, as in the body of the paper, with: $m_F T_{F,t} + m_n T_{n,t} = 0$.

For the market to clear, investors must be willing to deviate from the above target portfolios conditional on price. In Proposition 1, the fast investors' holdings deviate from their target portfolio as a linear function of price. Using that same functional form here, investors' holdings (conditional on price) are given by downward sloping functions (of price). Note that we are using a lower case p to denote the permanent component of price:

$$\begin{aligned} \pi_{n,t} &= T_{n,t} - a_n p_t \\ \pi_{F,t} &= T_{F,t} - a_F p_t \end{aligned}$$

Then, by market clearing,

$$\begin{aligned} m_n \pi_{n,t} + m_F \pi_{F,t} &= \\ m_n I_t - (a_n m_n + a_F m_F) p_t &= 0 \\ \Rightarrow \quad p_t &= \frac{m_n}{a_n m_n + a_F m_F} I_t \end{aligned}$$

Thus, market clearing implies that if optimal holdings are linear in price, then price is linear in I_t. The permanent price with dividends in body of the paper is $p_t^p = \sigma_w B_t$. Given that B_t and I_t are both Brownian motions, setting $\sigma_w = \frac{m_n}{a_n m_n + a_F m_F}$ yields an equivalent permanent price process due to dividends or imbalanced shocks: $p_t^p = p_t$.

46 We use the term "permanent price dynamics" as shorthand for the price process purged of all pricing errors (i.e., it is the accumulation of permanent price changes).

Additionally, using the same equivalence between B_t and I_t and assuming B_t and I_t both have the same correlation ρ with Z_t yields:

$$\mathrm{Corr}(\mathrm{d}Z_t,\mathrm{d}B_t)=\mathrm{Corr}\left(\mathrm{d}Z_t,\mathrm{d}p_t^p\right)=\rho=\mathrm{Corr}(\mathrm{d}Z_t,\mathrm{d}p_t)=\mathrm{Corr}(\mathrm{d}Z_t,\mathrm{d}I_t).$$

Thus, imbalanced shocks can yield a permanent price process that is equivalent to the with-dividend permanent price process in the body of the paper. In addition, if the correlation between the imbalanced and balanced shocks is the same as the correlation between the balanced shocks and the dividend process in the main paper, then the correlation between the balanced shocks and the changes in the 'imbalanced' permanent price process and the 'dividend' permanent price process is the same. Overall, our example illustrates how imbalanced shocks can generate correlations between the balanced shock process and returns as found in the body of the paper.

It is important to note that the above intuition relies on the assumption that the permanent price process is linear in the imbalanced shocks. While the temporary price process is a linear function as shown in Proposition 1, it is unlikely the permanent price would be linear in equilibrium. Price being linear allows us to use the Kalman filter to compute the exact likelihood function when structurally estimating the model. Hence, in the main text we solve the model with ρ, which has linear prices as opposed to the model with imbalanced shocks.

D. Limited-Attention Model and Its Equilibrium

D.1. Equilibrium Structure

Our approach to solve for an equilibrium is to "guess and verify." Concretely, we assume a certain functional form, or ansatz, for the equilibrium price process of the risky asset. Several parameters in the ansatz are assumed to be known by the agents but are, at first, left unspecified. In Section D.2. (below) we solve for the optimal individual policies taking the ansatz as given. In Section D.3. we pin down the unspecified parameters by imposing market clearing. If a choice of parameters in the ansatz allows the equilibrium conditions to hold, the ansatz is shown to be ex post rational.

We turn to a description of our ansatz. The equilibrium behavior of our economy is driven by the "gap process." We make three assumptions regarding the equilibrium structure. Existence and uniqueness of equilibrium, shown below, then prove these assumptions to be rational.

First, we assume that the gap process follows a multidimensional Ornstein-Uhlenbeck process.

Assumption 1 (Ornstein-Uhlenbeck). The dynamics of the gap process is

$$\mathrm{d}G_t=-\Lambda G_t\mathrm{d}t+\sigma_G\mathrm{d}Z_t \tag{D.1}$$

for a mean-reversion speed $\Lambda\in\mathbb{R}^{N\times N}$ and a diffusion matrix $\sigma_G\in\mathbb{R}^{N\times N}$.

In equilibrium, the mean-reversion speed in (D.1) is the diagonal matrix of attention intensities. This is shown in the proof of Proposition 2 below. To avoid verbosity we, however, already use the notation Λ for the mean-reversion speed.

Second, we assume that the price P_t of the risky asset is linear in the components of the gap process.

Assumption 2 (Linear Equilibrium). The price of the risky asset satisfies

$$P_t:=-p^\top G_t \tag{D.2}$$

for a vector $p\in\mathbb{R}^N$.

Finally, we assume the following for the information structure.

Assumption 3 (Gap is public information). All investors know the current value of the gap process when they make portfolio decisions.

As our analysis focuses on risk sharing, it is natural to abstract from other economic mechanisms, including asymmetric information. Assumption 3 makes all investors have the same expectations regarding risky returns. Without Assumption 3 investors would have to filter the current value of the gap process and investors in different classes would reach different estimates. This filtering would only obscure the risk-sharing mechanisms.

D.2. Individual Problems

In this subsection, we characterize the optimal policies of all investors conditional on the assumptions of Section D.1. regarding the equilibrium structure.

Fast investors. A fast investor i chooses its policy at t to solve

$$\max_{C,\pi} \mathrm{E}_t \left[\int_t^{\infty} e^{-r(u-t)} \left(\mathrm{d}C_u - \frac{r\gamma_F \sigma_w^2}{2} \left(T_{F,u} - \pi_{i,u}\right)^2 \mathrm{d}u \right) \right], \tag{D.3}$$

with admissible strategies (C,π) satisfying three conditions. First, the consumption C, the risky holdings π, and the wealth w satisfy the *budget constraint*:

$$\mathrm{d}w_t = r w_t \mathrm{d}t - \mathrm{d}C_t + \pi_t (\mathrm{d}D_t + \mathrm{d}P_t - r P_t \mathrm{d}t). \tag{D.4}$$

Second, to prevent infinite financing of consumption with debt, the *no-Ponzi condition*

$$\lim_{T\to\infty} e^{-r(T-t)} \mathrm{E}_t(w_T) = 0 \tag{D.5}$$

must hold for any time $t > 0$. Third, to prevent so-called doubling strategies, the *regularity condition*

$$\mathrm{E}_t \left(\int_t^T \pi_s^2 \mathrm{d}s \right) < +\infty \tag{D.6}$$

holds for any $t < T$. Finally, the expectation $\mathrm{E}_t[\cdot]$ in (D.3) is conditional on the current target portfolio $T_{I,t}$ and wealth w_t of the institution, along with the current value G_t of the gap process.

Market makers. A market maker i chooses its policy at t to solve

$$\max_{C,\pi} \mathrm{E}_t \left[\int_t^{\infty} e^{-r(u-t)} \left(\mathrm{d}C_u + - \frac{r\gamma_M \sigma_w^2}{2} \left(\pi_{i,u}\right)^2 \mathrm{d}u \right) \right]. \tag{D.7}$$

Just like fast investors, a policy (C,π) is admissible for a market maker if it satisfies the budget constraint (D.4), the no-Ponzi condition (D.5), and the regularity condition (D.6).

Lemma 1 (Holdings). We have the following three expressions:

- A market maker holds $\pi_{M,t}$ shares of the risky asset where $p \in \mathbb{R}^N$ and I_N is the identity matrix in $\mathbb{R}^{N\times N}$:

$$\pi_{M,t} = \frac{1}{r\gamma_M \sigma_w^2} \left[\frac{1}{\mathrm{d}t} \mathrm{E}_t(\mathrm{d}P_t) - r P_t \right] = \frac{1}{r\gamma_M \sigma_w^2} \left[p^\top (r I_N + \Lambda) G_t \right] \tag{D.8}$$

- A fast institution holds $\pi_{F,t}$ shares of the risky asset:

$$\pi_{F,t} = T_{F,t} + \frac{1}{r\gamma_F \sigma_w^2} \left[p^\top (r I_N + \Lambda) G_t \right] \tag{D.9}$$

- A slow investor of class j who arrives at the market at time t holds $\pi_{j,t}$ shares:

$$\pi_{j,t} = T_{j,t}. \tag{D.10}$$

Proof of Lemma 1. A time $t=0$ a fast investor i maximizes

$$\mathrm{E}_0\left[\int_0^{+\infty} e^{-ru}\left(\mathrm{d}\tilde{C}_u - \frac{r\gamma_F\sigma_w^2}{2}\left(T_{F,u}-\tilde{\pi}_{i,u}\right)^2\mathrm{d}u\right)\right] \tag{D.11}$$

over the admissible strategies $(\tilde{C},\tilde{\pi})$. A strategy is admissible if it satisfies a budget constraint, a no-Ponzi condition, and a certain regularity condition.[47]

By combining the budget constraint,

$$\mathrm{d}\tilde{w}_u = r\tilde{w}_u\mathrm{d}u - \mathrm{d}\tilde{C}_u + \tilde{\pi}_{i,u}(\mathrm{d}P_u - rP_u\mathrm{d}u), \tag{D.12}$$

and Itô's product rule, we can rewrite the discounted incremental consumption as

$$e^{-ru}\mathrm{d}\tilde{C}_u = e^{-ru}\tilde{\pi}_u(\mathrm{d}P_u - rP_u\mathrm{d}u) - \mathrm{d}\left(e^{-ru}\tilde{w}_u\right). \tag{D.13}$$

Then, by using the no-Ponzi,

$$\lim_{T\to\infty}\mathrm{E}\left[e^{-rT}\tilde{w}_T\right]=0, \tag{D.14}$$

and by injecting (D.13) into (D.11), we can rewrite the objective function of our investor as

$$\begin{aligned}
&\mathrm{E}_0\left[\int_0^{+\infty} e^{-ru}\left(\mathrm{d}\tilde{C}_u - \frac{r\gamma_F\sigma_w^2}{2}\left(T_{F,u}-\tilde{\pi}_{i,u}\right)^2\mathrm{d}u\right)\right]\\
&= w_0 + \mathrm{E}_0\left[\int_0^{+\infty} e^{-ru}\left(\tilde{\pi}_{i,u}(\mathrm{d}P_u - rP_u\mathrm{d}u) - \frac{r\gamma_F\sigma_w^2}{2}\left(T_{F,u}-\tilde{\pi}_{i,u}\right)^2\mathrm{d}u\right)\right]\\
&= w_0 + \mathrm{E}_0\left[\int_0^{+\infty} e^{-ru}\left(\tilde{\pi}_{i,u}\mathrm{E}_u[\mathrm{d}P_u - rP_u\mathrm{d}u] - \frac{r\gamma_F\sigma_w^2}{2}\left(T_{F,u}-\tilde{\pi}_{i,u}\right)^2\mathrm{d}u\right)\right]\\
&= w_0 + \mathrm{E}_0\left[\int_0^{+\infty} e^{-ru}\left(\tilde{\pi}_{i,u}p^\top(rI_n+\Lambda)G_u - \frac{r\gamma_F\sigma_w^2}{2}\left(T_{F,u}-\tilde{\pi}_{i,u}\right)^2\mathrm{d}u\right)\right],
\end{aligned} \tag{D.15}$$

where we used the law of iterated expectations for the second equality and the Assumptions 1 and 2 for the third equality. In particular, any admissible consumption plan $\tilde{C}$ is equally good for our investor.

Let us now consider the unique pointwise maximizer $\pi_{i,u}$ of the term between the parentheses in the last line of (D.15):

$$\pi_{i,u} = T_{F,u} + \frac{1}{r\gamma_F\sigma_w^2}p^\top(rI_N+\Lambda)G_u. \tag{D.16}$$

As inspection shows, this unique maximizer defines an admissible strategy and, as measured by (D.11), no other admissible strategy is better than the strategy defined by (D.16). In particular, (D.16) is the optimal trading strategy for our investor, as stated in the proposition.

The argument for a market maker is identical, up to the target portfolio being 0 at all times. □

D.3. Equilibrium: Holdings and Price

Our equilibrium definition is standard and combines individual optimality with market clearing for the risky asset.

47 The use of a tilde ($\tilde{}$) denotes *any* given admissible strategy (e.g., $\tilde{\pi}$ as any admissible trading strategy), whereas the notation without a tilde denotes its *optimal* counterpart (e.g., π).

Definition 1 (Equilibrium). An equilibrium consists of policies $\{\pi_i\}_{i\in\{I,M\}\cup\mathbf{N}}$ giving the risky holdings of all classes of investors, parameters Λ and σ_G defining the dynamics of the gap process as in Assumption 1, and a vector p defining the price of the risky asset as in Assumption 2.

These quantities satisfy three conditions.

(i) *Individual optimality*: The policies $\{\pi_i\}_{i\in\{I,M\}\cup\mathbf{N}}$ are given by Proposition 1 with the parameters of the gap process being set at their equilibrium values.

(ii) *Market clearing:*

$$m_F\pi_{F,t}+m_M\pi_{M,t}+\mathbf{1}_{(1\times N)}\mathrm{diag}(m_1,\dots,m_N)\pi_{\mathbf{N},t}\equiv 0, \tag{D.17}$$

with $\pi_{\mathbf{N},t}:=\left(\pi_{i,t}\right)_{i=1}^N$.

Lemma 2 (Price). An equilibrium exists and is unique. The gap process follows an Ornstein-Uhlenbeck process:

$$\mathrm{d}G_t=-\Lambda G_t\mathrm{d}t+\mathrm{diag}(\mu_1,\dots,\mu_N)\mathrm{d}Z_t, \tag{D.18}$$

where $\mu_j:=m_j\sigma_j$ is the total "risk mass" of investors in class j. The equilibrium price of the risky asset is

$$P_t=-p^\top G_t \quad \text{with} \quad p^\top=\frac{\sigma_w^2}{\frac{m_F}{r\gamma_F}+\frac{m_M}{r\gamma_M}}\mathbf{1}_{(1\times N)}(rI_N+\Lambda)^{-1}. \tag{D.19}$$

The equilibrium price process determines the dynamics of the trading policies shown in Proposition 1 with the row vector of weights, $p^\top$, being explicit in (D.19).

Proof of Lemma 2. The inelastic demand of all slow investors, the definition of the gap process, and a heuristic application of a cross-sectional strong law of large numbers (SLLN) yields the dynamics

$$\mathrm{d}G_t=-\Lambda G_t\mathrm{d}t+\mathrm{diag}(\mu_1,\dots,\mu_N)\mathrm{d}Z_t \tag{D.20}$$

for the gap process.[48]

We must still make sure that the assumed price process

$$P_t=-p^\top G_t \tag{D.21}$$

is consistent with market clearing. Concretely, market clearing at time t amounts to

$$m_M\pi_{M,t}+m_F\pi_{F,t}+\mathbf{1}_{(1\times N)}A_t=0\Leftrightarrow\left(\left(\frac{m_M}{r\gamma_M\sigma_w^2}+\frac{m_F}{r\gamma_F\sigma_w^2}\right)p^\top(rI_N+\Lambda)-\mathbf{1}_{(1\times N)}\right)G_t=0, \tag{D.22}$$

where we used Proposition 1, the definition $T_{F,t}$ in (6), and the definition of G_t in (9). As market clearing holds at any point in time and any state of the world in equilibrium, the price sensitivities p must be

$$p^\top=\frac{\sigma_w^2}{\frac{m_M}{r\gamma_M}+\frac{m_F}{r\gamma_F}}\mathbf{1}_{(1\times N)}(rI_N+\Lambda)^{-1}, \tag{D.23}$$

as stated in the proposition. □

48 See, for example, Judd (1985) and Sun (2006) for rigorous discussions of cross-sectional SLLNs.

D.4. Proof of Proposition 1

We combine the results from Lemma 1 and its associated proof in Section D.2. with the results from Lemma 2 and its associated proof in Section D.3. The result proves Proposition 1 from the main paper. □

E. Model-Implied Discrete-Time Dynamics

In this appendix, we provide additional details on the components of the model implied dynamics given in (20). The model moments calculated in Section 2.1 can be obtained using the full model dynamics by setting $\rho=0$ and using the scalar λ in place of the matrix Λ. The VAR in (20) includes the coefficient matrix V discussed and given in the main text in (21). This incorporates the autoregressive component of the model dynamics.

Below, we provide details related to the shocks. This includes the variance and covariance of the shocks, as well as the coefficient matrix W that maps the shocks into the model's variables. Finally, we also provide analysis on the pricing errors and returns for the full model.

The coefficient matrix W is

$$W=\begin{array}{c} \\ 3 \\ 3 \\ 1 \\ 1 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} 3 & 3 & 1 & 3 & 3 \end{array} \\ \left[\begin{array}{ccccc} I_3 & 0 & 0 & 0 & 0 \\ 0 & I_3 & 0 & 0 & 0 \\ \beta_M \mathbf{1}_{(1\times 3)} & \beta_M \mathbf{1}_{(1\times 3)} & 0 & 0 & 0 \\ 0 & -\mathbf{1}_{(1\times 3)} & 0 & 0 & \mathbf{1}_{(1\times 3)} \\ -\beta_w \mathbf{1}_{(1\times 3)} B_i & -\beta_w \mathbf{1}_{(1\times 3)} B_r & 1 & 0 & 0 \end{array}\right] \end{array} \in \mathbb{R}^{9\times 13}, \tag{E.1}$$

with $B_j=\left(rI_3+\Lambda_j\right)^{-1}$ and $j\in\{i,r\}$. The intuition for the elements in W immediately follows from the discussion of the corresponding elements in V (see previous paragraphs). The only difference is that W pertains to new shocks in target portfolios (and not to changes in legacy inefficient holdings).

The error term is

$$\varepsilon_t=\left(\varepsilon_{1,t} \quad \varepsilon_{2,t} \quad \varepsilon_{3,t}\right)^\top, \tag{E.2}$$

where $\varepsilon_{1,t}\in\mathbb{R}^6$ captures the net change in the gap process and $\varepsilon_{3,t}\in\mathbb{R}^6$ captures the change in target portfolios. Note that these are highly correlated, but not identical. Only part of the target portfolio change enters the gap process because of intraperiod trading. $\varepsilon_{2,t}$ captures the dividend shock. The covariance matrix of ε_t is

$$\begin{aligned}
\operatorname{Var}\left(\varepsilon_{1,t}\right) &= \int_0^{\Delta t} e^{-\Lambda(\Delta t-u)}\Big(\overbrace{\left(1-\rho^2\right)\operatorname{diag}^2(\mu)}^{\substack{\text{Idiosyncratic}\\ \text{shocks}}}+\overbrace{\rho^2\mu\mu^\top}^{\substack{\text{Common}\\ \text{(shared)}\\ \text{shocks}}}\Big)\left(e^{-\Lambda(\Delta t-u)}\right)^\top \mathrm{d}u, && \text{(E.3)}\\
\operatorname{Var}\left(\varepsilon_{2,t}\right) &= \sigma_w^2\Delta t,\\
\operatorname{Var}\left(\varepsilon_{3,t}\right) &= \left(\left(1-\rho^2\right)\operatorname{diag}^2(\mu)+\rho^2\mu\mu^\top\right)\Delta t,\\
\operatorname{Cov}\left(\varepsilon_{1,t},\varepsilon_{2,t}\right) &= \int_0^{\Delta t} e^{-\Lambda(\Delta t-u)}\rho\mu\sigma_w \mathrm{d}u,\\
\operatorname{Cov}\left(\varepsilon_{1,t},\varepsilon_{3,t}\right) &= \int_0^{\Delta t} e^{-\Lambda(\Delta t-u)}\left(\left(1-\rho^2\right)\operatorname{diag}^2(\mu)+\rho^2\mu\mu^\top\right)\mathrm{d}u,\\
\operatorname{Cov}\left(\varepsilon_{2,t},\varepsilon_{3,t}\right) &= \rho\sigma_w\mu^\top\Delta t,
\end{aligned}$$

where ρ is the correlation between the dividend shock and infrequent investors' target portfolio shocks.[49]

Let $\tilde{V}$ be the submatrix of V consisting of only the nonzero columns:

$$\tilde{V} = V_{(.,1:6)}, \tag{E.4}$$

where the subscript indicates the rows and columns that are selected with a dot used to select them all (above, $\tilde{V}$ has selected all the rows and the first six columns of V). Then (20) can then be written as

$$Y_t = \tilde{V} G_{t-\Delta t} + W \varepsilon_t. \tag{E.5}$$

Model-implied covariance and autocovariance matrices. The variance matrix for the gap process G_t is easily derived by taking Δt to infinity for $\mathrm{Var}\left(\varepsilon_{1,t}\right)$ in (E.3). The result is twice the variance of the gap process (i.e., $\mathrm{Var}(G_t - G_{t-\Delta t}) = \mathrm{Var}(G_t) + \mathrm{Var}(G_{t-\Delta t}) - 2\mathrm{Cov}(G_t, G_{t-\Delta t})$ where the last terms vanishes for $\Delta t \uparrow \infty$). Let

$$\mathrm{Var}(G_t) = \Big(\overbrace{\left(1-\rho^2\right)\mathrm{diag}^2(\mu)}^{\substack{\text{Idiosyncratic}\\ \text{shocks}}} + \overbrace{\rho^2 \mu\mu^\top}^{\substack{\text{Common}\\ \text{(shared)}\\ \text{shocks}}} \Big) \circ \left((\Lambda_{kk} + \Lambda_{ll})^{-1} \right)_{0 \le k,l \le 6} \tag{E.6}$$

where $\circ$ is the Hadamard product (i.e., element-wise product).

The simple first-order autoregressive structure of Y_t implies that its variance is

$$\mathrm{Var}(Y_t) = \tilde{V}\,\mathrm{Var}(G_t)\,\tilde{V}^\top + W\,\mathrm{Var}(\varepsilon_t)\,W^\top \tag{E.7}$$

and its autocovariance of order $n > 0$ is

$$\mathrm{Cov}(Y_t, Y_{t-n}) = \tilde{V} e^{-(n-1)\Lambda\Delta t} \begin{pmatrix} \mathrm{Var}(G_t) \\ \mathrm{Cov}(G_t, MMInv_t) \\ \mathrm{Cov}(G_t, RetFlow_t) \\ \mathrm{Cov}(G_t, Return_t) \end{pmatrix}^\top. \tag{E.8}$$

The autocovariance function in (E.8) shows that, not surprisingly, all decay is governed by the individual gap components (in $e^{-(n-1)\Lambda\Delta t}$). The decay could, however, still be different for different variable pairs,[50] as they load differently on the gap components (governed by $\tilde{V}$).

Characterizing pricing errors and returns. The model-implied variance and autocovariances can be used to develop several additional results that generate further economic insight. One particularly useful result is that the estimation delivers a full characterization of the pricing errors, their size, a decomposition, and their decay. To generate these results, let us first define the pricing error at time t as

$$s_t = W_{(9,1:6)} G_t \in \mathbb{R}. \tag{E.9}$$

Its variance therefore is

$$\mathrm{Var}(s_t) = W_{(9,1:6)}\,\mathrm{Var}(G_t)\,W_{(9,1:6)}^\top. \tag{E.10}$$

The structure of $\mathrm{Var}(G_t)$ admits a decomposition of pricing error variance into idiosyncratic components associated with the various slow-investor classes and frequencies and a common

---

[49] Note that element (i,j) from $\mathrm{Var}\left(\varepsilon_{1,t}\right)$ is $\frac{\mu_i^2}{2\Lambda_{ii}}\left(1-e^{-2\Lambda_{ii}\Delta t}\right)$ for $i=j$. This element is $\frac{\rho^2\mu_i\mu_j}{\Lambda_{ii}+\Lambda_{jj}}\left(1-e^{-\left(\Lambda_{ii}+\Lambda_{jj}\right)\Delta t}\right)$ for $i \neq j$.

[50] For example, one could compare the decay (for increasing n) of $(MMInv_t, RetFlow_{t-n})$ and $(MMInv_t, Return_{t-n})$.

factor correlated with the fundamental-value shock. Such a decomposition immediately follows from the structure of $\mathrm{Var}(G_t)$ in (E.6). The autocorrelation function for pricing errors (s_t) that defines their decay follows from the autocovariance function of G_t:

$$\rho_{s,1} = \frac{W_{(9,.)} e^{-\Lambda \Delta t} \mathrm{Var}(G_t) W_{(9,.)}^{\top}}{W_{(9,.)} \mathrm{Var}(G_t) W_{(9,.)}^{\top}}. \tag{E.11}$$

An alternative way to characterize the persistence of pricing errors is to compute their half-life. This is most easily done by looking at the bottom panel in Figure 3. The downward line crosses the 0.5 level after 31 (trading) days, or 6.2 weeks.

Another useful result is to compute the extent to which returns are "polluted" by pricing errors. More specifically, we want to know how different components of the pricing errors affect returns. The variance of returns immediately follows from (E.7):

$$\mathrm{Var}(r_t) = \overbrace{\tilde{V}_{(9,.)} \mathrm{Var}(G_t) \tilde{V}_{(9,.)}^{\top}}^{\text{Legacy error reduction}} + \overbrace{W_{(9,.)} \mathrm{Var}(\varepsilon) W_{(9,.)}^{\top}}^{\text{New shocks}}, \tag{E.12}$$

where the "new shocks" component can be further decomposed into

- A fundamental-value change component corresponding to $\mathrm{Var}(\varepsilon)_{(7,7)}$,
- New idiosyncratic target shock components corresponding to the diagonal of $\mathrm{Var}(\varepsilon)_{(1:6,1:6)}$, and a
- New shared component (correlated with fundamental-value change) corresponding to all off-diagonals of $\mathrm{Var}(\varepsilon)_{(1:7,1:7)}$.

The return autocorrelations immediately follow from (E.8) as the risky asset's return is the last element of Y_t (see (19)).

References

Abel, A. B., J. C. Eberly, and S. Panageas. 2007. Inaction and adjustment: Consequences for households and firms. *American Economic Review Papers & Proceedings* 97:244–9.

———. 2013. Optimal inattention to the stock market with information costs and transactions. *Econometrica* 81:1455–81.

Afonso, G. and R. Lagos. 2015. Trade dynamics in the market for federal funds. *Econometrica* 83:263–313.

Almgren, R. and N. Chriss. 2001. Optimal execution of portfolio transactions. *Journal of Risk* 3:5–39.

Asparouhova, E., H. Bessembinder, and I. Kalcheva. 2010. Liquidity biases in asset pricing tests. *Journal of Financial Economics* 96:215–37.

———. 2013. Noisy prices and inference regarding returns. *Journal of Finance* 68:665–714.

Bacchetta, P. and E. van Wincoop. 2010. Infrequent portfolio decisions: A solution to the forward discount puzzle. *American Economic Review* 100:870–904.

Bao, J., J. Pan, and J. Wang. 2011. The illiquidity of corporate bonds. *Journal of Finance* 66:911–46.

Bertsimas, D. and A. W. Lo. 1998. Optimal control of execution costs. *Journal of Financial Markets* 1:1–50.

Biais, B. 1993. Price formation and equilibrium liquidity in fragmented and centralized markets. *Journal of Finance* 48:157–85.

Biais, B., J. Hombert, and P.-O. Weill. 2014. Equilibrium pricing and trading volume under preference uncertainty. *Review of Economic Studies* 81:1401–37.

Bogousslavsky, V. 2016. Infrequent rebalancing, return autocorrelation, and seasonality. *Journal of Finance* 71:2967–3006.

Brennan, M. J. and A. W. Wang. 2010. The mispricing return premium. *Review of Financial Studies* 23:3437–68.

Campbell, J. Y., S. J. Grossman, and J. Wang. 1993. Trading volume and serial correlation in stocks returns. *The Quarterly Journal of Economics* 108:905–39.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay. 1997. *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Cella, C., A. Ellul, and M. Giannetti. 2013. Investors horizons and the amplification of market shocks. *Review of Financial Studies* 26:1607–48.

Chien, Y. L., H. Cole, and H. Lustig. 2012. Is the volatility of the market price of risk due to intermittent portfolio rebalancing? *American Economic Review* 102:2859–96.

Cochrane, J. H. 1994. Permanent and transitory components of GNP and stock prices. *Quarterly Journal of Economics* 109:241–65.

Crouzet, N., I. Dew-Becker, and C. G. Nathanson. 2019. On the effects of restricting short-term investment. *Review of Financial Studies* 33:1–43.

Duffie, D. 2010. Presidential address: Asset price dynamics with slow-moving capital. *Journal of Finance* 65:1237–67.

Duffie, D., N. Gârleanu, and L. H. Pedersen. 2007. Valuation in over-the-counter markets. *Review of Financial Studies* 20:1865–900.

Durbin, J. and S. J. Koopman. 2012. *Time series analysis by state space models*. Oxford, UK: Oxford University Press.

Fama, E. F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25:383–417.

———. 1991. Efficient capital markets: II. *Journal of Finance* 46:1575–617.

Gârleanu, N. 2009. Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory* 144:532–64.

Griffin, J. M., J. H. Harris, and S. Topaloglu. 2003. The dynamics of institutional and individual trading. *Journal of Finance* 58:2285–320.

Grossman, S. J. and M. H. Miller. 1988. Liquidity and market structure. *Journal of Finance* 43:617–33.

Hasbrouck, J. and G. Sofianos. 1993. The trades of market makers: An empirical analysis of NYSE specialists. *Journal of Finance* 48:1565–93.

Hendershott, T. and A. J. Menkveld. 2014. Price pressures. *Journal of Financial Economics* 114:405–23.

Hendershott, T. and P. C. Moulton. 2011. Automation, speed, and stock market quality: The NYSE's hybrid. *Journal of Financial Markets* 14:568–604.

Hendershott, T. and M. S. Seasholes. 2007. Market maker inventories and stock prices. *American Economic Review* 97:210–14.

Hu, X., J. Pan, and J. Wang. 2013. Noise as information for illiquidity. *Journal of Finance* 68:2341–82.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *Journal of Finance* 45:881–98.

Judd, K. L. 1985. The law of large numbers with a continuum of IID random variables. *Journal of Economic Theory* 35:19–25.

Kaniel, R., G. Saar, and S. Titman. 2008. Individual investor trading and stock returns. *Journal of Finance* 63:273–310.

Koijen, R. S. J. and M. Yogo. 2019. A demand system approach to asset pricing. *Journal of Political Economy* 127:1475–515.

Lagos, R. and G. Rocheteau. 2009. Liquidity in asset markets With search frictions. *Econometrica* 77:403–26.

Lakonishok, J., A. Shleifer, and R. W. Vishny. 1992. The impact of institutional trading on stock prices. *Journal of Financial Economics* 32:23–43.

Lehmann, B. 1990. Fads, martingales, and market efficiency. *Quarterly Journal of Economics* 105:1–28.

Llorente, G., R. Michaely, G. Saar, and J. Wang. 2002. Dynamic volume-return relation of individual stocks. *Review of Financial Studies* 15:1005–47.

Lo, A. W., H. Mamaysky, and J. Wang. 2004. Asset prices and trading volume under fixed transaction costs. *Journal of Political Economy* 112:1054–90.

Madhavan, A. and S. Smidt. 1993. An analysis of changes in specialist inventories and quotations. *Journal of Finance* 48:1595–1628.

Nagel, S. 2012. Evaporating liquidity. *Review of Financial Studies* 25:2005–2039.

Nieuwerburgh, S. Van and L. Veldkamp. 2009. Information immobility and the home bias puzzle. *Journal of Finance* 64:1187–15.

Nofsinger, J. R. and R. W. Sias. 1999. Herding and feedback trading by institutional and individual investors. *Journal of Finance* 46:2263–95.

Poterba, J. M. and L. H. Summers. 1988. Mean reversion in stock prices evidence and implications. *Journal of Financial Economics* 22:27–59.

Roll, R. 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* 39:1127–39.

Sims, C. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50:665–690.

Sun, Y. 2006. The exact law of large numbers via Fubini extension and characterization of insurable risks. *Journal of Economic Theory* 126:31–69.

Weller, B. M. 2018. Does algorithmic trading reduce information acquisition? *Review of Financial Studies* 31:2184–2226.