Asset Price Dynamics with Limited Attention

Terrence Hendershott  
U.C. Berkeley  

Albert J. Menkveld  
VU University Amsterdam  

Rémy Praz  
Copenhagen Business School  

Mark S. Seasholes  
Arizona State  

This Version: 15-Jan-2018

Abstract

Inattention can significantly distort security prices. We develop a structural model with stochastic investor inattention and private value shocks that net to zero across investors. We solve the model analytically and estimate its parameters with NYSE data. The model is able to explain the empirical relations among market-maker inventories, retail order flows, and stock returns from daily to monthly frequencies. The structural estimation further identifies pricing errors. These errors distort prices by 2.9 percentage points and have a half-life of 3.0 weeks. The pricing errors account for 27% of daily, and 19% of monthly, idiosyncratic return variances.

Keywords: Transitory Volatility, Limited Attention, Market Makers

JEL Nums: G12, G14
Inattention can significantly distort security prices. We develop a structural model with stochastic investor inattention and private value shocks that net to zero across investors. We solve the model analytically and estimate its parameters with NYSE data. The model is able to explain the empirical relations among market-maker inventories, retail order flows, and stock returns from daily to monthly frequencies. The structural estimation further identifies pricing errors. These errors distort prices by 2.9 percentage points and have a half-life of 3.0 weeks. The pricing errors account for 27% of daily, and 19% of monthly, idiosyncratic return variances.
How much do observable stock prices deviate from fundamental values? And when they do, how long do these “pricing errors” last? Financial economists have long known that asynchronously arriving (or inattentive) investors could be the root cause of these errors. The errors compensate market makers who supply liquidity by stepping in to match buyers and sellers across time.

The pricing errors induced by liquidity supply cause price changes to show a pattern of predictable reversals. Short-run reversals are a major focus of the market microstructure literature, while lower-frequency reversals are typically studied in the asset pricing literature. Our paper attempts to link these literatures by studying the magnitude of pricing errors in typical New York Stock Exchange (NYSE) stocks at frequencies from a day to a month.

Our primary contribution is to analytically solve a structural model with inattention and estimate it by fitting the dynamics of NYSE market-maker inventories, retail order flows, and prices. Retail flows serve as a proxy for trading by some of the inattentive investors. We estimate the model using daily data to fit variances and autocorrelations (including the cross-autocorrelations) with lags up to a month. The fit of the estimated model is encouraging as all these dynamic relations are matched, both in terms of sign and magnitude. This is particularly noteworthy since the model is only a single friction away from a standard asset-pricing model. The pricing errors it identifies are economically meaningful. They account for 27% of daily and 19% of monthly idiosyncratic return variances.

Existing, related models have not been structurally estimated possibly due to a lack of analytical solutions or inadequate data. Instead, empirical work has focused on rare events where large supply shocks help identify links between persistent pricing errors and investors’ inattention (Duffie, 2010). Our approach is quite different and systematically identifies continuously arriving hedging shocks that cause pricing errors. The result is the ability to measure the magnitude and duration of pricing errors for typical NYSE stocks on typical days.

Our Theoretical Approach: Our theoretical model is recursive in nature, assumes that all 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. Hendershott and Seasholes (2007) and Hendershott and Menkveld (2014) link these reversals to market maker inventories. Our paper differs from the earlier work by considering how market makers interact with other groups of investors (both attentive and inattentive).
are atomistic price-takers, and runs in continuous time. The model’s core distinguishing feature is the inclusion of multiple groups of inattentive investors who operate at different frequencies. Such inattention is the only friction in the model (i.e., information is symmetric and agents are zero-mass price-takers). The private-value shocks that investors experience perfectly offset one another. Absent the inattention friction, trade is purely re-allocational, does not require intermediation, and does not affect prices (i.e., no pricing errors). However, if at least one group of investors is inattentive the model predicts non-trivial price and trade patterns.

The channels that generate these patterns are perhaps best explained by considering an example. 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. These investors will endogenously become market makers. Divide the remaining mass in half and let the private-value shocks that one half experiences be perfectly 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, 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 pulls in market makers who purchase the securities that the fast investors want to sell. It also induces these fast investors to reduce their liquidity demands and spread them through time. Both the fast investors and the market makers will sell to slow investors once the latter group 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 groups, and the inattention frequency of the slow investors together determine the magnitude and duration of the pricing errors.

Having only one friction (inattention) both clarifies the channels at work in the model and sets up a challenging data fitting exercise. Modeling the inventory control choices of market makers is highlighted in 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 in the market for liquidity provision.
Table 1

Stock return autocorrelations at various frequencies

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 equation (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 701 U.S. stocks. In parentheses below each of our autocorrelations is a t-statistic testing if the correlation is zero. The model-implied autocorrelations are based on estimates presented in Section 3.

<table>
<thead>
<tr>
<th></th>
<th>Period</th>
<th>Daily</th>
<th>Monthly</th>
<th>Bi-monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campbell, Lo, &amp; MacKinlay</td>
<td>1962-1994</td>
<td>-0.03</td>
<td>-0.04</td>
<td></td>
</tr>
<tr>
<td>Our data</td>
<td>1998-2005</td>
<td>-0.02</td>
<td>-0.05</td>
<td>(-8.1)</td>
</tr>
<tr>
<td>Model-implied</td>
<td>1998-2005</td>
<td>-0.01</td>
<td>-0.06</td>
<td></td>
</tr>
</tbody>
</table>

From Theory to Structural Estimation: The model’s strength is its simplicity and versatility. Our model is invariant to the sampling frequency. Section 3 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, or a month. Our model, therefore, can be used by both monthly asset pricers and sub-millisecond microstructure researchers. In addition, allowing for multiple classes of slow investors who operate at different frequencies turns out to be a crucial feature when fitting NYSE price and trading dynamics. In particular, the autocorrelation in daily idiosyncratic returns decays too slowly to be explained using only a single class of slow investors who arrive stochastically at an average daily frequency. We find a significantly better fit using two classes of slow investors: one with investors who arrive semi-daily (twice a day) and another with investors who arrive monthly.

The monthly inattentive investors lead to the slowly decaying pricing errors found in the autocorrelations of NYSE returns. The presence of this class is also the main reason for why pricing errors are sizeable. We estimate that prices deviate from fundamental values by 2.9 percentage points with a half-life of three weeks.

---

5 All returns in the paper are idiosyncratic. Hereafter, to ease exposition we typically refer to them simply as “returns.”
6 Section 3 shows that adding a third group of investors does not improve the model’s fit.
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 bi-monthly frequency than at a monthly frequency for their 1962-1994 sample. In our 1998-2005 sample, we find a similar pattern when comparing daily with monthly returns. The table further shows that our model can produce such a pattern. The model-implied autocorrelations are more negative for monthly returns than for daily returns.

The intuition for why such patterns can occur is best developed by taking pricing-error persistence to a limit. At high frequencies such errors will wash out when taking first differences (i.e., when computing stock returns) and the first-order autocorrelation will tend to zero. At low frequencies the errors will decay enough to “contaminate” returns and cause a negative first-order autocorrelation. The argument is developed more formally in Appendix A. This insight should caution researchers not to conclude that prices are “efficient” when seeing negligible first-order autocorrelation in returns sampled at high frequency.\footnote{Papers estimating both a stock’s fundamental price and its pricing error are Poterba and Summers (1988), Roll (1988), and Cochrane (1994). Papers such as Poterba and Summers (1988) assume that pricing errors evolve as a specific autoregressive process while we use NYSE trade data to identify the timing and dynamics of pricing errors. Put differently, it is very difficult to separate pricing errors from fundamental values using only observable prices in a finite sample. Using trade data helps avoid such difficulties.}

**Novel Results from the Structural Estimation:** The structural model estimation yields novel insights in four, broad areas. First, the model requires two groups of slow investors in order to achieve a reasonable fit: semi-daily slow investors and monthly slow investors. These categories feature both slow institutions and retail investors. Institutions are more prevalent at both the semi-daily and the monthly frequencies.\footnote{Because 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 groups.} While retail investors are a small part of the market, they make up a relatively larger part of monthly investors.\footnote{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 monthly slow investors than of semi-daily slow investors.} These observations are based on estimates of the total mass of...
private-value shocks, referred to as “risk mass.” 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 and the hedge shocks they experience, we are able to identify the product of the two.

Second, the model fit improves significantly when including a “leverage effect.” Innovations to fundamental values are estimated to be positively correlated with shocks to fast institutions’ target portfolios. Therefore, a large drop in a stock’s fundamental value implies a reduction in the target holdings of fast institutions and, by the mechanism of the model, a negative shock to the pricing error. One explanation for this could be fast institutions being more highly levered than the slow investors. Being constrained by such leverage, fast institutions might have to sell when the stock’s fundamental value drops. The resulting amplification of shocks is related to work by Gromb and Vayanos (2002) and Brunnermeier and Pedersen (2009).

Third, the risk-bearing capacity of all fast participants (i.e., market makers plus fast institutions) is also identified. A $1 million shock to fast institutions’ target portfolios leads to a pricing error of only -5 basis points.

Fourth, the model allows for a decomposition of the pricing error variance. The standard deviation of the various components are: 2.7% due to monthly slow investors, 0.6% due to the leverage effect, and 0.1% due to semi-daily slow investors. A further decomposition reveals that 3.9% of pricing error variance is due to slow retail investors while 96.1% is due to slow institutional investors. The structural estimation allows deeper identification of the source of pricing errors than has previously been possible.

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

\footnote{Chien, Cole, and Lustig (2012) explore how inattention in the form of intermittent rebalancing increases the volatility of the market price of risk.}
classes of inattentive (slow) investors. Importantly, our inattentive investors are atomistic and arrive stochastically. This feature keeps the dimension of the state space small enough to make structural estimation feasible and thus identification of pricing errors possible (a detailed argument is at the end of Section 1.3). Our closed-form solutions also allow us to decompose the pricing errors into easily understood economic quantities.

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 market wide 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.

Pricing errors (or noise) have more recently been linked to biases in asset pricing tests. Our paper provides estimates of the magnitude and duration of noise which could then be used to measure biases following, for example, Asparouhova, Bessembinder, and Kalcheva (2010) and Asparouhova, Bessembinder, and Kalcheva (2013). In a related paper, Brennan and Wang (2010) show that pricing errors affect a stock’s expected return beyond what can be attributed to fundamental risk.

1 Asset pricing model with limited attention

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, 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. Evaluating a GMM penalty function is trivial for each point in the parameter space, which turns out to be a sine qua non for fitting the model to the data (i.e., we have 10 parameters to fit 54 moments, see Section 3). Third, our model becomes invariant to the sampling frequency. Section 2 shows how one can convert the implied model dynamics from $dt$ to $\Delta t$ 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 sub-millisecond microstructure researchers.

**Assets.** There are two assets in the economy. First, there is a risky asset in zero net supply that pays dividends over any interval \((t, t + dt]\), with \(B\) being a Brownian motion.

\[
dD_t = \sigma_w dB_t
\]

Adding a positive expected dividend or a positive net supply of the risky asset affects the fundamental value of the asset but not the deviations of the actual price from its fundamental value. As we are interested in these deviations, it is convenient to set this fundamental value to zero at all times. The dividend dynamics of (1) serve this aim.

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 \(\mathbb{N} := \{1, \ldots, N\}\) denote all classes of slow investors. We index all of the \(N + 2\) classes with \(j \in \{F, M\} \cup \mathbb{N}\). There is a continuum of atomistic agents in each of the \(N + 2\) classes. The masses of the investor classes are \(m_F, m_M, m_1, \ldots, m_N\), 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 \(\lambda_i\) and the Poisson processes are independent across investors (even within a class).\(^{11}\) For convenience, we define \(\Lambda\) to be the diagonal matrix whose entries are the attention intensities of the slow investors.

\[
\Lambda := \text{diag} \left( \lambda_1, \ldots, \lambda_N \right)
\]

\(^{11}\) A more general setting would allow for correlation in attention 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 towards 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 in a few paragraphs). Concretely, at time $t$, an investor $i$ of class $j$
chooses his policies to maximize

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

(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,
Gahrleanu, and Pedersen (2007), Gahrleanu (2009), Lagos and Rocheteau (2009), and Afonso and Lagos

To ensure that the model’s full dynamics become available in closed-form we assume slow investors
are infinitely risk-averse (i.e., $\gamma_j = \infty, \forall j \in \mathbb{N}$).\footnote{Nagel (2012) uses myopia to facilitate closed-form solutions.} This is a technical assumption that removes speculation by slow classes. We believe the assumption is natural as it fits the behavior of inattentive
investors. People who trade monthly are unlikely to provide liquidity. Also, as the frequency between trades becomes longer, the cost of speculative trading (quadratic loss) becomes greater.

Target portfolios. An $N$-dimensional Brownian motion $Z$ drives the slow-investors’ target portfo-
ilios. 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.

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

(4)
The target portfolio of the market makers is zero at all times and shown in (5). This definition is consistent with market makers only trading to facilitate risk-sharing among the other market participants.

\[ T_{M,t} := 0. \]  

(5)

Finally, the (scalar) target portfolio of the fast institutions is shown in (6) where \( \mathbf{1}_{(k \times 1)} \) is a \( k \times 1 \) matrix of ones.\(^{13}\)

\[ T_{F,t} := -\frac{1}{m_F} \mathbf{1}_{(1 \times N)} \text{diag}(m_1, \ldots, m_N) T_{N,t}. \]  

(6)

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:

\[ \sum_{j \in \{F, M\} \cup N} m_j T_{j,t} = 0. \]  

(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., zero in our setting).

The Brownian motions (\( B \) driving the dividend process and \( Z \) driving the target portfolios) are allowed to be correlated. Assuming this correlation to be \( \rho \) yields:

\[ \text{Corr} (dZ_t, dB_t) = \rho \mathbf{1}_{N \times 1}. \]  

(8)

A non-zero correlation between the fundamentals and a target portfolio can allow for wealth effects in reduced form. In particular, it is natural to imagine that investors may wish to reduce their risk exposure (in absolute terms) after seeing their wealth eroded by bad returns. While intuitive, such an effect cannot be generated by CARA investors nor by our CARA-like investors. A correlation between the target shocks and dividends can, however, capture such wealth effects in a reduced-form manner. Finally, note that adding a non-zero correlation between fundamentals and target portfolios is consistent with our micro-foundations and the preferences in (3) remain unchanged. This is intuitive as investors are risk-neutral regarding shocks to their target portfolios. They do not adjust their portfolio strategies when these shocks become correlated with fundamentals.

\(^{13}\)In our model, we abstract away from target shocks affecting the risky asset’s fundamental value. The assumption is made explicit here: the fast institutions’ target portfolio is equal and opposite to a weighted sum of the slow investors’ targets.
The model estimates (presented in Section 3.3) suggest that we refer to a non-zero correlation as a “leverage effect.” We find that the target shocks of fast institutions are positively correlated with dividend shocks. We interpret this finding as consistent with the fast institutions use of leverage—fast institutions want to unwind some of their speculative positions when there are negative fundamental shocks.

The gap process (state variable). Finally, it is useful 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 (atomistic) slow investors in the $N$ different classes. More precisely,

$$G_t := \text{diag}(m_1, \ldots, m_N) (T_{N,t} - A_{N,t}) \in \mathbb{R}^N,$$

(9)

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

$$A_{i,t} := \int_{u \in m_i} \pi_{u,t} du.$$

(10)

This gap process turns out to be the state variable upon which all the model’s dynamics depend. Defining the gap process at an investor-class level benefits from the independent 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 investor 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 Individual problems

Our approach to solve 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. A more detailed description is in Appendix C.1. This appendix also discusses the three key assumptions/guesses used:
the gap process is OU, the pricing errors are linear in the gap process, and the gap process is public
information (labeled Assumptions 1, 2, and 3, respectively in the appendix).  

The optimal holdings of all agents are available in closed form and are stated in the following
proposition (with a proof in Appendix C.2).

**Proposition 1.** A market maker optimally holds

\[
\pi_{M,t} = \frac{1}{r\gamma_M\sigma_w^2} \left[ \frac{1}{dt} E_t (dP_t) - rP_t \right] = \frac{1}{r\gamma_M\sigma_w^2} \left[ p^\top (rI_N + \Lambda) G_t \right]
\]  

(11)

shares of the risky asset where \( p \in \mathbb{R}^N \) and \( I_N \) is the identity matrix in \( \mathbb{R}^{N\times N} \). A fast institution optimally holds

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

(12)

shares of the risky asset. If he arrives at the market at time \( t \), a slow investor of class \( j \) optimally chooses

\[
\pi_{j,t} = T_{j,t}.
\]  

(13)

Proposition 1 shows how hedging and speculative motives define the optimal portfolios. The fast
institution’s portfolio in (12) features both motives additively. The first RHS term involves his target
portfolio and therefore represents hedging. The second RHS term loads positively on a weighted sum
of the gap process with weights proportional to the row vector \( p^\top \). In the equilibrium result that
follows (below), 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 investor makes a speculative
profit in expectation. Note further that he holds more (than the target portfolio) when he is less
risk-averse (\( \gamma_F \)) or when the asset has less fundamental risk (\( \sigma_w \)).

Finally, note that the market maker’s portfolio in (11) features only the speculative motive as his
target portfolio is zero. The slow investor’s portfolio in (13) only features a hedging motive as, by
assumption, he does not engage in speculation.

\[\text{14 As agents are risk-neutral in terms of consumption with a time preference parameter equal to the interest rate, any policy in which consumption eventually takes place is equally good. Therefore, no consumption policy is reported. Note that delaying consumption forever is not optimal.}\]
1.3 Equilibrium

We turn to the main equilibrium result. It is based on a standard definition of equilibrium that features individual optimality, market clearing, and rational expectations (see Appendix C.3 for this definition and a proof of the following proposition).

**Proposition 2 (Equilibrium).** Equilibrium exists and is unique. The gap process is an Ornstein-Uhlenbeck one:

\[ dG_t = -\Lambda G_t dt + \text{diag}(\mu_1, \ldots, \mu_N) dZ_t, \quad (14) \]

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

\[ P_t = -p^\top G_t \quad \text{with} \quad p^\top = \frac{\sigma_w^2}{rF} + \frac{mM}{rM} 1_{1 \times N} (rI_N + \Lambda)^{-1}. \quad (15) \]

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 (15).

Proposition 2 leads to the following observations. First, the OU process for the gap vector in (14) is intuitive. It shows that class \( j \)'s gap decays smoothly with intensity \( \lambda_j \). The independent arrivals of atomistic investors generate the smoothness. The size of gap shocks scales with the mass of investors in this class \( (m_j) \) times the size of an individual-investor shock \( (\sigma_j) \).

Second, the price impact row vector \( p^\top \) that translates portfolio-holding gaps to pricing errors yields several insights. Higher fundamental risk \( (\sigma_w) \) or lower risk absorption capacity of fast investors increase 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.\(^{15}\)

Third, combining the results of Propositions 1 and 2 yields 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

\(^{15}\)The larger premium for lower interest rates and, at the same time, less discounting (see preferences in (3)) is more challenging to explain. It appears that temporarily tying up capital in speculative positions is more expensive in our economy.
fundamental risk increases in equilibrium to the point that fast investors willingly take it on (i.e., note that the $\sigma^2_w$ in (12) cancels against $\sigma^2_w$ in (15)). 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., $N = 4$ in Section 3.3’s final model—shown in the last column of Table 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. Compare our stochastic-arrivals set-up to 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 20. An additional benefit of our model is that it yields analytic expressions for any dimensionality while Duffie’s model generally does not.

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 $\Delta t$ periods. The section first derives the model dynamics to provide expressions for variances, covariances, and autocorrelations. In Section 3, we use GMM to match the moments to those from NYSE data. Our data’s sampling period is one day. To keep the structural estimation numerically tractable we limit the limited-attention (slow) investors to be of three types:

- Type $s$ investors who, on average, arrive at the market semi-daily, say in the morning and in the afternoon,
- Type $d$ investors who, on average, arrive once a day, and
- Type $m$ investors who, on average, arrive once a month.

As our data is daily, we pick one type ($d$) to be this frequency. We then choose a faster type ($s$) and a slower type ($m$). These types arrive at the market with Poisson intensities such that average
durations are twice a day for $s$, once a day for $d$, or once a month for $m$. The matrix with Poisson intensities therefore is:

$$
\Lambda_j = \text{diag} (\lambda_s, \lambda_d, \lambda_m) = \text{diag} \left( 2, 1, \frac{1}{20} \right), \quad j \in \{i, r\}.
$$  \hfill (16)

For each type, we assume that slow investors are either institutional (“$i$”) or retail (“$r$”). The reason for this further categorization is that we have retail flow data.\(^{16}\) The investor-type subscripts are thus $\{s, d, m\} \times \{i, r\}$. The gap processes for the $3 \times 2$ types of slow investors become:

$$
G_{j,t} = \begin{pmatrix} G_{sj,t} & G_{dj,t} & G_{mj,t} \end{pmatrix}^\top \in \mathbb{R}^3, \quad j \in \{i, r\}.
$$  \hfill (17)

These gaps are associated with investors that collectively have risk masses (i.e., $\mu = m\sigma$):

$$
\mu_j = \begin{pmatrix} \mu_{sj} & \mu_{dj} & \mu_{mj} \end{pmatrix}^\top \in \mathbb{R}^3, \quad j \in \{i, r\}.
$$  \hfill (18)

For each of these variables, it is convenient to also define a companion variable that stacks both types of slow investors. Let us denote such variables by dropping the subscripts. For example, $\Lambda$ becomes a $6 \times 6$ matrix with $\Lambda_i$ and $\Lambda_r$ on the (block) diagonal. $G_t$ and $\mu$ are defined analogously as $6 \times 1$ column vectors.

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

$$
Y_t = \begin{pmatrix} G_{i,t}^\top & G_{r,t}^\top & \text{MMInv}_t & \text{RetFlow}_t & \text{Return}_t \end{pmatrix}^\top \in \mathbb{R}^9
$$  \hfill (19)

where $\text{MMInv}_t, \text{RetFlow}_t, \text{Return}_t \in \mathbb{R}$ 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:

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

The coefficient matrices and variables are as follows. The coefficient matrix $V$ is (with row and column dimension in lightgray along the axes):

\(^{16}\)We refer to individuals as “retail” so we can use different single-letter subscripts for institutions and individuals.
\[
V = \begin{pmatrix}
\begin{bmatrix} I_3 & 0 & 0 \end{bmatrix} & \begin{bmatrix} e^{-\Lambda_i \Delta t} & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} \\
\begin{bmatrix} 0 & e^{-\Lambda_i \Delta t} & 0 \end{bmatrix} & \begin{bmatrix} \beta_M^1 & 0 & 0 \end{bmatrix} & \begin{bmatrix} \beta_M^2 \end{bmatrix} & \begin{bmatrix} \beta_M^3 \end{bmatrix} & \begin{bmatrix} \beta_M^4 \end{bmatrix} & \begin{bmatrix} \beta_M^5 \end{bmatrix} \\
\begin{bmatrix} 0 & I_3 - e^{-\Lambda_i \Delta t} & 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} \\
\begin{bmatrix} 1 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} & \begin{bmatrix} 1 \end{bmatrix} & \begin{bmatrix} \beta_w^1 \end{bmatrix} & \begin{bmatrix} \beta_w^2 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} \beta_w^3 \end{bmatrix} & \begin{bmatrix} \beta_w^4 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{bmatrix} 0 \end{bmatrix} & \begin{b}
change enters the gap process because of intra-period trading. \( \varepsilon_{2,t} \) captures the dividend shock. The covariance matrix of \( \varepsilon_t \) is:

\[
\begin{align*}
\text{Var} (\varepsilon_{1,t}) &= \int_0^{\Delta t} e^{-\Lambda(\Delta t-u)} \left( (1 - \rho^2) \text{diag}^2 (\mu) + \rho^2 \mu \mu^\top \right) \left( e^{-\Lambda(\Delta t-u)} \right)^\top \, du, \\
\text{Var} (\varepsilon_{2,t}) &= \sigma_w^2 \Delta t, \\
\text{Var} (\varepsilon_{3,t}) &= \left( (1 - \rho^2) \text{diag}^2 (\mu) + \rho^2 \mu \mu^\top \right) \Delta t, \\
\text{Cov} (\varepsilon_{1,t}, \varepsilon_{2,t}) &= \int_0^{\Delta t} e^{-\Lambda(\Delta t-u)} \rho \mu \sigma_w \, du, \\
\text{Cov} (\varepsilon_{1,t}, \varepsilon_{3,t}) &= \int_0^{\Delta t} e^{-\Lambda(\Delta t-u)} \left( (1 - \rho^2) \text{diag}^2 (\mu) + \rho^2 \mu \mu^\top \right) \, du, \\
\text{Cov} (\varepsilon_{2,t}, \varepsilon_{3,t}) &= \rho \sigma_w \mu^\top \Delta t,
\end{align*}
\]

where \( \rho \) is the correlation between the dividend shock and infrequent investors’ target portfolio shocks.\(^{17}\)

Let \( \tilde{V} \) be the sub-matrix of \( V \) consisting of only the non-zero columns:

\[
\tilde{V} = V_{(.,1:6)},
\]

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:

\[
Y_t = \tilde{V} G_{t-\Delta t} + W \varepsilon_t.
\]

**Model-implied covariance and auto-covariance matrices.** The variance matrix for the gap process \( G_t \) is easily derived by taking \( \Delta t \) to infinity for \( \text{Var} (\varepsilon_{1,t}) \) in (25). The result is twice the variance of the gap process (i.e., \( \text{Var} (G_t - G_{t-\Delta t}) = \text{Var} (G_t) + \text{Var} (G_{t-\Delta t}) - 2 \text{Cov} (G_t, G_{t-\Delta t}) \)) where

\(^{17}\)Note that element \((i,j)\) from \( \text{Var} (\varepsilon_{1,t}) \) is \( \frac{\sigma^2}{\xi_i^2} \left( 1 - e^{-2\Lambda_i \Delta t} \right) \) for \( i = j \). This element is \( \frac{\rho^2 \mu_i \mu_j}{\Lambda_i + \Lambda_j} \left( 1 - e^{-(\Lambda_i + \Lambda_j) \Delta t} \right) \) for \( i \neq j \).
the last terms vanishes for $\Delta t \uparrow \infty$). Let

\[
\text{Var} (G_t) = \left( (1 - \rho^2) \, \text{diag}^2 (\mu) + \rho^2 \mu \mu^\top \right) \circ \left( (\Lambda_{kk} + \Lambda_{ll})^{-1} \right)_{0 \leq k, l \leq 6}
\]

(28)

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:

\[
\text{Var} (Y_t) = \tilde{V} \text{Var} (G_t) \tilde{V}^\top + W \text{Var} (\varepsilon_t) W^\top
\]

(29)

and its auto-variance of order $n \geq 0$ is:

\[
\text{Cov} (Y_t, Y_{t-n}) = \tilde{V} e^{-(n-1)\Lambda \Delta t} \begin{pmatrix}
\text{Var} (G_t) \\
\text{Cov} (G_t, \text{MMInv}_t) \\
\text{Cov} (G_t, \text{RetFlow}_t) \\
\text{Cov} (G_t, \text{Return}_t)
\end{pmatrix}^\top.
\]

(30)

The autocovariance function in (30) 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,\(^{18}\) as they load differently on the gap components (governed by $\tilde{V}$).

**Characterizing pricing errors and returns.** The model-implied variance and auto-covariances 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:

\[
s_t = W_{(9,1:6)} G_t \in \mathbb{R}.
\]

(31)

Its variance therefore is:

\[
\text{Var} (s_t) = W_{(9,1:6)} \text{Var} (G_t) W_{(9,1:6)}^\top.
\]

(32)

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

\(^{18}\text{For example, one could compare the decay (for increasing } n \text{) of (MMInv}_t, \text{RetFlow}_{t-n}) \text{ and (MMInv}_t, \text{Return}_{t-n}).\)
with the fundamental-value shock. Such a decomposition immediately follows from the structure of \( \text{Var}(G_t) \) in (28). The autocorrelation function for pricing errors \( (s_t) \) that defines their decay follows from the autocovariance function of \( G_t \):

\[
\rho_{s,1} = \frac{W_{(9,.)} e^{-\Lambda \Delta t} \text{Var}(G_t) W_{(9,.)}^T}{W_{(9,.)} \text{Var}(G_t) W_{(9,.)}^T}.
\]  

(33)

An alternative way to characterize the persistence of pricing errors is to compute their half life. For a (multivariate) one standard deviation shock to the error, the half life \( h \) is computed by solving the following equation:

\[
1_{(1 \times 6)} \left( e^{-\Lambda h} - \frac{1}{2} I_6 \right) \mu = 0.
\]  

(34)

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 follows immediately from (29):

\[
\text{Var}(r_t) = \underbrace{\hat{V}_{(9,.)} \text{Var}(G_t) \hat{V}_{(9,.)}^T}_{\text{Legacy error reduction}} + \underbrace{W_{(9,.)} \text{Var}(\varepsilon) W_{(9,.)}^T}_{\text{New shocks}}.
\]  

(35)

where the “new shocks” component can be further decomposed into:

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

The return autocorrelations follow immediately from (30) as the risky asset’s return is the last element of \( Y_t \) (see (19)).
3 Estimation

3.1 Data

Our data start in January 1999 and end in December 2005 and comes from four datasets:

- 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.\(^{19}\)

- 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 percent 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, e.g., hedge funds

\(^{19}\)The investor-type classifications 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.
and, more recently, high frequency traders, compete with the specialist by placing limit orders to supply liquidity.\textsuperscript{20}

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.

If our retail trading and market maker inventory data are representative of all such market participants, then all of our estimation results remain unchanged except for two. Both the market-makers’ share of the total risk-bearing capacity and the relative size of retail investors in the total population of slow-investors become lower bounds.\textsuperscript{21}

We start with the 2,357 common stocks that can be matched across the NYSE, TAQ, and CRSP databases. We construct a balanced panel of data to ensure results are comparable throughout time. 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 701 actively traded stocks.\textsuperscript{22}

\textbf{Idiosyncratic variables} Risks associated with market-wide return shocks can be hedged with highly-liquid index products. Therefore, they need not cause any pricing errors for individual stocks. For this reason the empirical analysis focuses on the idiosyncratic components of our variables.

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 $\text{MMInv}_t$ to denote the

\textsuperscript{20}Hendershott and Moulton (2011) show the NYSE’s market structure changes after our sample period (2006-7) lead to a reduced role for the specialist and a decline in the NYSE’s share of trading. These evolutions highlight 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. It is unclear when or if such data may become available.

\textsuperscript{21}The respective lower bounds pertain to estimates of $\beta_M$ and Figure 6 in Section 3.3. We convert market makers’ inventory positions and retail investor net trades to US 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.

\textsuperscript{22}The sample is, with the exception of four stocks, the same as the one used in Hendershott and Menkveld (2014). For a more detailed characterization of the stocks, please see that paper.
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 are 97.4% of total variance for $\text{MMInv}_t$ and 99.9% for $\text{RetFlow}_t$. 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 non-systematic effects; order flows and positions due to (market-wide) systematic effects are removed by the procedure described above.

Figure 1 plots all autocorrelations for our three variables 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 nine plots at the top show the (cross) autocorrelations of all possible ordered pairs of the three series. Note that contemporaneous correlations are shown at lag zero for the off-diagonal plots. The bottom row consists of three bar charts to illustrate the variance of each series. These 12 plots illustrate the multivariate auto-covariance function for all series with lags ranging from zero days to 20 days (monthly). The plots summarize the dynamics of our data series.

The figure has some notable and statistically significant patterns. First, the standard deviation of market maker inventories is $0.7$ million. Inventories initially decay rather quickly as the first-day autocorrelation is 0.5. After a day, they decay slowly and end with a 20-day autocorrelation of 0.2. Second, the standard deviation in retail flows is $1.1$ million. Similar to market maker inventories, they decay extremely quickly on the first day (even more quickly than do market maker inventories) and then slowly in the following nineteen days. Third, the standard deviation in idiosyncratic daily returns is 2.5%. The return autocorrelations are significantly negative throughout the first twenty days suggesting that at least part of the original pricing error is persistent.

The figure further reveals strong cross autocorrelations. First, market maker inventories and retail flows are positively correlated, both contemporaneously and through time (see plots (2,1) and (1,2) in the figure). This pattern is consistent with our model. Periods when market makers are long

---

23The plots show the cross-sectional average moments along with their confidence intervals based on a cross-section of the 701 stocks. We first compute the variance and (cross) autocorrelations for each stock. The observations for all 701 stocks are then collected into a cross-sectional sample. This sample is then used to do statistical inference (i.e., compute means, standard errors, and error correlations).
Figure 1. Empirical moments. These graphs plot the empirical moments with a 95% confidence bound.
securities correspond with periods of retail investors buying. This 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 (0.01). 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 (see plot (3,2)). 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. This behavior benefits individuals as a group as they all seem to buy below fundamental values. Even those who arrive late seem to buy at depressed prices—to see this point, one has to cumulate a strong contemporaneous negative return with modest future positive returns (see plot (2,3)).

3.2 Matching the moments

A standard general method of moments (GMM) is used to estimate the model’s deep parameters. The model-implied moments are matched to the empirical moments discussed in the previous subsection (i.e., Figure 1). Specifically, we focus on the observed variances of the three variables: $MMInv_t$, $RetFlow_t$, and $Return_t$ as well as contemporaneous correlations and lagged (cross) autocorrelations. To make estimation feasible, we set up our GMM in the following way:

- All moments are weighted equally. The GMM penalty function therefore sums the squared differences between the empirical moments and the model-implied moments for a particular value of the deep parameters. To ensure that the GMM penalty function does not focus on some moments disproportionately, we chose the data units such that the means (of volatility)—depicted in the bottom-row plots of Figure 1—are the same order of magnitude as the correlations that are matched (i.e., they are close to the (-1,1) interval).

- Instead of matching all autocorrelations and cross autocorrelations for all 20 lags as depicted in
Figure 1, we bin the data by week to reduce the dimensionality of the optimization and keep estimation feasible. We nevertheless separate out the first-order autocorrelations as these exhibit strong patterns that would be lost if averaged with the other days in the first week. So, in total we match \((4 \text{ weeks} + 1 \text{ day}) \times (9 \text{ variables pairs}) + 3 \text{ standard deviations} + (6 \text{ contemporaneous correlations}) = 54 \text{ moments}\). Note that the contemporaneous correlations appear twice as the \(\text{Corr}(X_t, Y_t) = \text{Corr}(Y_t, X_t)\).

The optimization is done in two stages. To limit the possibility of becoming trapped in a local minimum, we first compute the GMM penalty for a grid that covers a reasonable range of parameter values. We then start a steepest-descent algorithm from the grid point with the lowest penalty value. The procedure finds the minimum in the immediate neighborhood of this point. The optimization is done in Python with help of the Pandas package and the parallel computing package Fire of Tange (2011).

3.3 Results

This section presents all estimation results. To illustrate how multiple classes of slow investors impact our model’s dynamics, we start by estimating parameters using a single frequency.\(^{24}\) For simplicity we set the inattention frequency equal to the sampling frequency, which is daily. Figure 2 illustrates the first model’s best possible fit. We refer to fits with differing numbers of slow investor classes as different models, although they are all based on the same underlying theoretical model. The dots in all plots are the target moments computed from the data. The lines in the top rows and the bars in the bottom row are the closest the model can get (i.e., the model-implied moments based on GMM estimates).

Figure 2 illustrates the first model’s poor fit in terms of dynamics. The curvature in the autocorrelation decays appear hard to match as is evident from the three diagonal plots in top nine graphs. For example, the model-implied market maker inventories initially drop too steeply. Later, the model-implied inventories flatten out too much and we are not able to establish a good fit throughout. The poor fits motivate us to use multiple frequencies of slow investors. In particular, the inability to match

\(^{24}\)The model and estimation always include fast institutions and market makers.
Parameters best fit: \((\mu_{si}, \mu_{di}, \mu_{mi}) = (0.0, 29.8, 0.0), (\mu_{sr}, \mu_{dr}, \mu_{mr}) = (0.0, 1.8, 0.0), \beta_M = 0.032, \beta_w = 0.00013, \sigma_w = 0.023, \rho = -0.17\)

Figure 2. First model’s fit using only daily slow investors. Parameter estimates are from Table 2, 1st Model which only includes daily slow investors. The blue lines in the top three rows and blue shaded bars in the bottom row are the model-implied moments. The dark blue dots are the empirical moments. It is estimated using GMM and targeting all dots (moments) with equal weights.
Table 2
Parameter estimates

This table presents the GMM parameter estimates and their standard errors. Subscript “i” refers to slow institutions and “r” to retail investors. The table further contains the GMM criterion-function value to compare fit across columns. The GMM penalty value is expressed in units of 1/1000. The stars (*/**/***)) indicate statistical significance at a 10%, 5%, and 1% level, respectively.

<table>
<thead>
<tr>
<th>Inclusive frequencies</th>
<th>1st model</th>
<th>2nd model</th>
<th>3rd model</th>
<th>4th model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Daily</td>
<td>Daily and monthly</td>
<td>Semi-daily, daily, and monthly</td>
<td>Semi-daily and monthly</td>
</tr>
<tr>
<td><strong>Panel A: Mass of slow institutions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{si}$</td>
<td>0.65</td>
<td>65.0***</td>
<td>65.0***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.9)</td>
<td>(1.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{di}$</td>
<td>29.8***</td>
<td>39.0***</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.3)</td>
<td>(0.7)</td>
<td>(1.2)</td>
<td></td>
</tr>
<tr>
<td>$\mu_{mi}$</td>
<td>6.6***</td>
<td>9.0***</td>
<td>9.0***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1)</td>
<td>(0.1)</td>
<td>(0.1)</td>
<td></td>
</tr>
<tr>
<td><strong>Panel B: Mass of (slow) retail investors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{sr}$</td>
<td></td>
<td>1.4***</td>
<td>1.4***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td></td>
</tr>
<tr>
<td>$\mu_{dr}$</td>
<td>1.8***</td>
<td>1.8***</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td></td>
</tr>
<tr>
<td>$\mu_{mr}$</td>
<td>0.9***</td>
<td>1.8***</td>
<td>1.8***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0)</td>
<td>(0.0)</td>
<td>(0.0)</td>
<td></td>
</tr>
<tr>
<td><strong>Panel C: Other parameters</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_M$</td>
<td>0.032***</td>
<td>0.019***</td>
<td>0.015***</td>
<td>0.015***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>$\beta_w$</td>
<td>0.00013***</td>
<td>0.00005***</td>
<td>0.00005***</td>
<td>0.00005***</td>
</tr>
<tr>
<td></td>
<td>(0.000011)</td>
<td>(0.000008)</td>
<td>(0.000009)</td>
<td>(0.000009)</td>
</tr>
<tr>
<td>$\sigma_w$</td>
<td>0.023***</td>
<td>0.022***</td>
<td>0.020***</td>
<td>0.020***</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.17***</td>
<td>-0.18***</td>
<td>-0.18***</td>
<td>-0.18***</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>GMM penalty</td>
<td>370</td>
<td>107</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>#stocks</td>
<td>701</td>
<td>701</td>
<td>701</td>
<td>701</td>
</tr>
</tbody>
</table>

The longer-lag autocorrelation suggests that using only daily inattention is insufficient to match the empirical data.

Table 2 presents the estimation results for four different empirical models. The models allow for different sets of slow investor frequencies. The first column of results contains the estimates using only daily frequency investors, as depicted in Figure 2. The second column of results adds slow investors who arrive at a monthly frequency, which helps to better fit the moments. The GMM penalty drops
by more than two-thirds from 0.000370 to 0.000107. The initial drop in retail flow autocorrelation however remains too steep in the data (as illustrated in Figure 7 in the online appendix). The third column of results therefore adds the semi-daily (i.e., twice daily) frequency and the fit again improves by almost two-thirds from 0.000107 to 0.000038.

The fit of the three-frequency version (shown in the third column of results) is notable in that all moments exhibit a reasonable match, both qualitatively and quantitatively. However, the masses at the daily frequency become statistically insignificant for both slow institutions and retail investors \((\mu_{di},\mu_{dr})\). Therefore, we drop the daily frequency in our fourth and final model. The fourth column of results contains this model’s estimates. The fit remains equally strong as the GMM penalty value is again 0.00038. Figure 3 illustrates this model’s fit. The fourth model’s fit is noticeably better than the fit in Figure 2 and is almost as good as the three-frequency model (in Figure 8 of the online appendix) while using only one less frequency of inattentive investors.

The parameter estimates of the final model are all statistically significant and lead to the following insights.\(^{25}\) First, the hedging mass of slow institutions is perhaps not surprisingly a lot higher than what it is for retail investors at both frequencies: 65.0 versus 1.4, respectively, for the semi-daily frequency and 9.0 versus 1.8 for the monthly frequency. Note further that comparing within institutions and retail, there is relatively more retail mass at the monthly frequency. This confirms our intuition that institutions, even the ones who are slow, are still relatively faster than retail investors. But, the results do show that there are some highly attentive retail investors who visit the market twice a day on average and some relatively inattentive institutions who visit the market once a month on average.

Second, the NYSE specialists absorb only 1.5% of trade imbalances that are caused by inattention. The other 98.5% is traced back to the fast institutions. The fast institutions either hold back on their desired trading and thus earn the liquidity premium (i.e., the pricing error) or some of them temporarily hold inefficient positions to earn this premium. Those that hold inefficient positions (deviations from their target portfolios) are participating in liquidity supply.

Third, the risk-bearing capacity of fast participants (market makers plus fast institutions) is sub-

\(^{25}\)Statistical inference in GMM involves the gradient \(\mathcal{G}\) of the model-implied moments \(\mathcal{B}\) with respect to the parameter vector \(\theta\). The covariance matrix of parameter estimates then is \((\mathcal{G}^\top \Omega^{-1} \mathcal{G})^{-1}\) where \(\Omega\) is the covariance matrix of empirical moments.
Parameters best fit: \((\mu_s, \mu_d, \mu_m) = (65.0, 0.0, 9.0), (\mu_s, \mu_d, \mu_m) = (1.4, 0.0, 1.8), \beta_M = 0.015, \beta_w = 4.8e-05, \sigma_w = 0.02, \rho = -0.18\)

Figure 3. Fourth model’s fit using semi-daily and monthly investors. This figure illustrates the fourth and final model’s fit. The blue lines in the top three rows and blue shaded bars in the bottom row are the model-implied moments. The dark blue dots are the empirical moments. It is estimated with GMM and targets all dots with equal weights. This model features slow investors who arrive at semi-daily and monthly frequencies.
A $1 million shock in the gap of fast investors times a factor that is approximately its duration (i.e., times \((r + \lambda)^{-1}\) with \(r \ll \lambda\)) commands a price pressure of -5 basis points. Note that such (inefficient) positions impose a \(\sigma_w=2\%\) daily standard deviation of fundamental risk on their holders.

Finally, note that shocks to target holdings of slow investors correlate negatively with fundamental-value changes as \(\rho = -0.18\). 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 in such conditions. On average, they cause a negative pricing error which adds to the price drop. We refer to the negative correlation as a leverage effect. We interpret it as fast investors becoming leverage-constrained and being forced to sell when fundamental values fall.

While the leverage label is purely speculative, some sort of amplification mechanism is needed to fit our model to the data. To illustrate this point, we force this correlation to be zero, re-optimize, and plot the fits in Figure 9 of the online appendix. Note that Figure 9 illustrates substantially poorer fits and the GMM penalty jumps from 0.00038 to 0.00100. The most revealing mismatch can be seen in the data’s strong negative contemporaneous correlation of retail flows and returns. This correlation cannot be matched by the \(\rho = 0\) model (see plot (3,2) or (2,3)). If \(\rho = 0\), then hedging shocks are uncorrelated with fundamental-value innovations which implies a zero contemporaneous correlation of retail flows and returns (which is counterfactual).

**Parameter identification.** How well are the deep parameters identified? We first empirically check whether the gradient of the model-implied moments with respect to the parameters is of full rank. Using parameter estimates from the fourth and final model, we find the gradient is of full rank. There are no numerical issues as this gradient matrix appears to be well conditioned and thus far from being perfectly multicollinear. The condition number of the standardized gradient is 32.7 and thus far from infinity.

The identification of the individual parameters often depends on a particular subset of moments.
Figure 4. 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., semi-daily and monthly). The bottom panel of the figure shows how pricing errors decay over time. These graphs are based on parameter estimates from the fourth and final model in Table 2.
For example, the shape of the decay in market maker inventories and pricing errors identifies the distribution of total hedging risk mass \( \mu \) across the various frequencies of slow investor groups (in our case, across semi-daily and monthly slow investors). The autocorrelation of retail flows identifies how much of this total mass is due to retail investors. The cross-autocorrelations of market-maker inventories and retail flows (i.e., plots (1,2) and (2,1) in Figure 3) determine the level of total hedging risk mass (as opposed to the distribution across frequencies). To see this, consider making retail flows a higher fraction of all slow investors’ flows. In this case, the cross-autocorrelations would become larger as more of the slow investors’ flows are observed (i.e., the slow institutions’ flows are unobserved and therefore add noise which depresses correlations).

With the total hedging demands (i.e., \( \mu \) vector) identified, the model determines the standard deviation of the total inefficient positions that the fast institutions need to absorb. Adding this to the standard deviation of market-maker inventories enables identification of the relative fraction of risk-absorption capacity that is from the market makers (\( \beta_M \)).

The price-trade cross-autocorrelations identify the remaining parameters. The response of pricing errors to market maker inventory shocks identifies the “price” of liquidity or \( \beta_w \). The non-autocorrelated part of return shocks identifies the size of fundamental-value innovations or \( \sigma_w \). Finally, a nonzero contemporaneous correlation between retail flows and returns identifies the leverage effect parameter \( \rho \). As discussed above, one can see the better fit by comparing the final-model fit in Figure 3 where \( \rho = -0.18 \), with Figure 9 in the online appendix, where \( \rho \) is constrained to be zero during the GMM estimation. If fundamental-value innovations were orthogonal to hedging shocks, then the contemporaneous correlation between returns and retail flows is predicted to be zero (compare plot (3,2) in both figures).

**Characterization of pricing errors.** The remainder of this section characterizes the pricing errors, their effect on returns, and how slow institutions and retail investors contribute to them. The upper panel of Figure 4 decomposes the steady-state pricing error variance into three components based on (32). The largest component is due to monthly slow investors and amounts to a standard deviation of 2.7%. Semi-daily slow investors only contribute 0.1% to the overall standard deviation. The difference in contributions is a stark contrast to the hedging masses of the two groups—the mass of semi-daily
Figure 5. Daily return variance decomposition. The graphs illustrate how total idiosyncratic return variance can be decomposed into a fundamental-value change component and three pricing error components: a legacy-error reduction component, a new target portfolio shocks’ component, and a leverage-effect component. The leverage effect is due to dividend shocks being negatively correlated with the slow investors’ target portfolio shocks. The upper panel illustrates the decomposition for daily return variance and the lower panel for monthly return variance. These graphs are based on parameter estimates from the fourth and final model.

The lower panel of Figure 4 illustrates the decay in pricing errors by plotting their autocorrelation function. The plot is based on (33). The plot shows that pricing errors are very persistent and decay slowly over time. After a full month (i.e., trading 20 days) a third remain. The reason is that the pricing errors are dominated by the monthly slow investors, as the decomposition in the upper panel clearly shows. The half-life of a pricing error shock is 3.0 weeks computed using (34).
Figure 6. Retail flow contribution to pricing errors. This figure illustrates how much retail flows contribute to the pricing error variance based on (idiosyncratic) target shocks. Slow investors consist of those arriving at semi-daily and monthly frequencies.

The persistence of pricing errors cause them to substantially affect both daily and monthly returns. Figure 5 illustrates this observation by decomposing returns into four components based on (35). Fundamental-value innovations constitute the largest component in both daily and monthly returns. Its standard deviation is 2.0% for daily returns and $\sqrt{20} \times 2.0\% = 9.0\%$ for monthly returns.

The standard deviations of the other components range from 0.2% to 0.9% for daily returns and from 1.8% to 3.1% for monthly returns. Note that the relative sizes of the legacy error reduction components represent the most salient difference between the daily and monthly return plots. The strong error persistence makes the legacy error reduction a small component in daily returns and a modest component in monthly 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 6 illustrates that retail flows contribute only a small part of the total pricing error variance. The contribution is 3.9% which is close to the squared ratio of the hedging mass of the monthly slow retail investors to the monthly slow institutional investors (i.e., $(1.8/9.0)^2$).

Overall, our structural estimation shows pricing errors are significant and long lasting.
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 trade data significantly improve the measurement of pricing errors’ role in stock return volatility. We find that pricing errors account for 27% of daily and 19% of monthly idiosyncratic return variances.

Our model and empirical approach can be applied to other data from a range of investor groups and over different 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 or private data providers such as Ancerno. The continuous-time model can also be translated into frequencies as high as a millisecond. Data from exchanges identifying high-frequency traders could therefore also be incorporated into our approach to examine these traders’ roles in correcting or possibly causing pricing errors.

In the future, it may also 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 could also 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 impact the duration of pricing errors. These extensions provide examples of potentially important future work.
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

$$\text{cov} (w_t + s_t - s_{t-1}, w_{t-1} + s_{t-1} - s_{t-2}) = -(1 + \rho_{s,2} - 2\rho_{s,1}) \sigma_s^2 < 0,$$  \hspace{1cm} (36)

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

$$\text{var} (w_t + s_t - s_{t-1}) = \sigma_w^2 + 2 (1 - \rho_{s,1}) \sigma_s^2.$$  \hspace{1cm} (37)

The first-order autocorrelation of daily returns is therefore

$$\rho_{r,1} = -\frac{(1 + \rho_{s,2} - 2\rho_{s,1}) \sigma_s^2}{\sigma_w^2 + 2 (1 - \rho_{s,1}) \sigma_s^2}.$$  \hspace{1cm} (38)

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 (38) becomes

$$\rho_{r,1} = -\frac{\sigma_s^2}{\sigma_w^2 + 2\sigma_s^2}.$$  \hspace{1cm} (39)

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 ($\sigma_s^2$ is small relative to $\sigma_w^2$), the above expression is small, negative, and approximately equal to minus the ratio of $\sigma_s^2$ to $\sigma_w^2$. 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 $\sigma_s^2$ is small relative to $\sigma_w^2$, the denominator in (39) 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:\(^{26}\)

\[
\lim_{\rho_{a,2} \uparrow 1} \rho_{r,1} = 0. \quad (40)
\]

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, 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 (38) 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.

\(^{26}\)Formally showing this limit requires additional assumptions about pricing error process, e.g., its variance must remain finite.
B  Notation summary

This appendix summarizes the notation used throughout our paper. We first describe the (measured) variables and then describe the estimated parameters.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t$</td>
<td>The period length implied by the sampling frequency.</td>
</tr>
<tr>
<td>$\varepsilon_t$</td>
<td>Vector with all the model’s shocks in the period from $t - \Delta t$ to $t$.</td>
</tr>
<tr>
<td>$G_t$</td>
<td>The gap between target and actual portfolio aggregated across all investors.</td>
</tr>
<tr>
<td>$\lambda_i$</td>
<td>Poisson arrival intensity investor type $i \in {s, d, m}$.</td>
</tr>
<tr>
<td>$\Lambda_k$</td>
<td>Diagonal matrix with the intensities of investor type $k \in {i, \rho}$ on the diagonal.</td>
</tr>
<tr>
<td>$m$</td>
<td>Mass of investors.</td>
</tr>
<tr>
<td>$\text{MMInv}_t$</td>
<td>Market makers’ inventory at time $t$.</td>
</tr>
<tr>
<td>$r$</td>
<td>Risk-free rate.</td>
</tr>
<tr>
<td>$\text{RetFlow}_t$</td>
<td>Retail investor order flow in the period from $t - \Delta t$ to $t$.</td>
</tr>
<tr>
<td>$\text{Return}_t$</td>
<td>Asset’s idiosyncratic (mid-quote) return in the period from $t - \Delta t$ to $t$.</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Per-investor shock in target portfolio.</td>
</tr>
<tr>
<td>$Y_t$</td>
<td>State vector that stacks all the model’s unobserved and observed variables.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The ten estimated parameters</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_w$</td>
<td>Asset’s fundamental-value risk relative to total risk-absorption capacity (i.e., market makers’ plus fast investors).</td>
</tr>
<tr>
<td>$\beta_M$</td>
<td>Market makers’ risk-absorption capacity relative to total capacity (i.e., market makers’ plus fast investors).</td>
</tr>
<tr>
<td>$\mu_{kl}$</td>
<td>The size of the total target portfolio shock aggregated across all slow investors of type $k \in {i, \rho}$ and frequency $l \in {s, d, m}$. Note 1: There are six $\mu_{kl}$ variables. Note 2: $\mu := m\sigma$ is the mass of investors times average per-investor shock size.</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Correlation of all slow investors target portfolio shocks and the asset’s fundamental-value change.</td>
</tr>
<tr>
<td>$\sigma_w$</td>
<td>Asset’s fundamental risk (i.e., standard deviation of the asset’s fundamental-value changes).</td>
</tr>
</tbody>
</table>
C Limited-attention model and its equilibrium

C.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 1.2 we solve for the optimal individual policies taking the ansatz as given. In Section 1.3 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 ex post to be 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 multi-dimensional Ornstein-Uhlenbeck process.

\textbf{Assumption 1 (Ornstein-Uhlenbeck).} The dynamics of the gap process is

\[ dG_t = -\Lambda G_t dt + \sigma_G dZ_t \]  

(41)

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 (41) 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 \( \Lambda \) 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.

\textbf{Assumption 2 (Linear Equilibrium).} The price of the risky asset satisfies

\[ P_t := -p^\top G_t \]  

(42)

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

Finally, we assume the following for the information structure.

\textbf{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.

C.2 Individual problems

In this subsection, we characterize the optimal policies of all investors conditional on the assumptions of Section C.1 regarding the equilibrium structure.
**Fast investors.** A fast investor $i$ chooses its policy at $t$ to solve
\[
\max_{C,\pi} \mathbb{E}_t \left[ \int_t^\infty e^{-r(u-t)} \left( dC_u - \frac{r\gamma F\sigma^2_w}{2} (T_{F,u} - \pi_{i,u})^2 \right) du \right],
\]
with admissible strategies $(C,\pi)$ satisfying three conditions. First, the consumption $C$, the risky holdings $\pi$, and the wealth $w$ satisfy the *budget constraint*:
\[
dw_t = rw_t dt - dC_t + \pi_t (dD_t + dP_t - rP_t dt).
\]
Second, to prevent infinite financing of consumption with debt, the *no-Ponzi condition*
\[
\lim_{T \to \infty} e^{-r(T-t)} \mathbb{E}_t (w_T) = 0
\]
must hold for any time $t > 0$. Third, to prevent so-called doubling strategies, the *regularity condition*
\[
\mathbb{E}_t \left( \int_t^T \pi^2_s ds \right) < +\infty
\]
holds for any $t < T$. Finally, the expectation $\mathbb{E}_t[\cdot]$ in (43) 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
\[
\max_{C,\pi} \mathbb{E}_t \left[ \int_t^\infty e^{-r(u-t)} \left( dC_u - \frac{r\gamma M\sigma^2_w}{2} (\pi_{i,u})^2 \right) du \right].
\]
Just like fast investors, a policy $(C,\pi)$ is admissible for a market maker if it satisfies the budget constraint (44), the no-Ponzi condition (45), and the regularity condition (46).

**Proof of Proposition 1.** A time $t = 0$ a fast investor $i$ maximizes
\[
\mathbb{E}_0 \left[ \int_0^{+\infty} e^{-ru} \left( d\tilde{C}_u - \frac{r\gamma F\sigma^2_w}{2} (T_{F,u} - \tilde{\pi}_{i,u})^2 \right) du \right]
\]
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.\(^\text{27}\)

By combining the budget constraint
\[
d\tilde{w}_u = r\tilde{w}_u du - d\tilde{C}_u + \tilde{\pi}_{i,u} (dP_u - rP_u du)
\]
and Itô’s product rule, we can rewrite the discounted incremental consumption as
\[
e^{-ru} d\tilde{C}_u = e^{-ru} \tilde{\pi}_{u} (dP_u - rP_u du) - d \left( e^{-ru} \tilde{w}_u \right).
\]
Then, by using the no-Ponzi
\[
\lim_{T \to \infty} \mathbb{E} \left[ e^{-rT} \tilde{w}_T \right] = 0,
\]
\(^\text{27}\) A $\tilde{\cdot}$ denotes any given admissible strategy (e.g., $\tilde{\pi}$ as any admissible trading strategy), whereas the notation without $\tilde{\cdot}$ denotes its *optimal* counterpart (e.g., $\pi$).
and by injecting (50) into (48), we can rewrite the objective function of our investor as

\[
E_0 \left[ \int_0^{+\infty} e^{-ru} \left( d\tilde{C}_u - \frac{r \gamma F \sigma^2_w}{2} (T_{F,u} - \tilde{\pi}_{i,u})^2 du \right) \right] = w_0 + E_0 \left[ \int_0^{+\infty} e^{-ru} \left( \tilde{\pi}_{i,u} (dP_u - rP_u du) - \frac{r \gamma F \sigma^2_w}{2} (T_{F,u} - \tilde{\pi}_{i,u})^2 du \right) \right] = w_0 + E_0 \left[ \int_0^{+\infty} e^{-ru} \left( \tilde{\pi}_{i,u} p^\top (rI_n + \Lambda) G_u - \frac{r \gamma F \sigma^2_w}{2} (T_{F,u} - \tilde{\pi}_{i,u})^2 du \right) \right],
\]

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 (52):

\[\pi_{i,u} = T_{F,u} + \frac{1}{r \gamma F \sigma^2_w} p^\top (rI_n + \Lambda) G_u.\]  

(53)

As inspection shows, this unique maximizer defines an admissible strategy and, as measured by (48), no other admissible strategy is better than the strategy defined by (53). In particular, (53) 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. \( \square \)

C.3 Equilibrium

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

**Definition 1** (Equilibrium). An equilibrium consists of policies \( \{\pi_i\}_{i \in \{I,M\} \cup N} \) giving the risky holdings of all classes of investors, parameters \( \Lambda \) and \( \sigma_G \) 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 \( \{\pi_i\}_{i \in \{I,M\} \cup N} \) are given by Proposition 1 with the parameters of the gap process being set at their equilibrium values.

(ii) **Market clearing:**

\[
m_F \pi_{F,t} + m_M \pi_{M,t} + 1_{(1 \times N)} \text{diag}(m_1, \ldots, m_N) \pi_{N,t} = 0,
\]

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

**Proof of Proposition 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

\[
dG_t = -\Lambda G_t dt + \text{diag}(\mu_1, \ldots, \mu_N) dZ_t
\]

(55)

for the gap process.\(^{28}\)

\(^{28}\)See, for example, Judd (1985) and Sun (2006) for rigorous discussions of cross-sectional SLLNs.
We must still make sure that the assumed price process

\[ P_t = -p^\top G_t \]  \hspace{1cm} (56)

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

\[ m_M \pi_{M,t} + m_F \pi_{F,t} + 1_{(1 \times N)} A_t = 0 \iff \left( \left( \frac{m_M}{r \gamma_M \sigma_w^2} + \frac{m_F}{r \gamma_F \sigma_w^2} \right) p^\top (r I_N + \Lambda) - 1_{(1 \times N)} \right) G_t = 0, \]  \hspace{1cm} (57)

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

\[ p^\top = \frac{\sigma_w^2}{\frac{m_M}{r \gamma_M} + \frac{m_F}{r \gamma_F}} 1_{(1 \times N)} (r I_N + \Lambda)^{-1}, \]  \hspace{1cm} (58)

as stated in the proposition. \( \square \)
References


Online Appendix to
“Asset Price Dynamics with Limited Attention”

This online appendix contains the following list of tables and figures references in the main document (note that to avoid confusion the numbering starts where the numbering of tables and figures in the main text left off).

List of Figures

7 Model fit, daily and monthly. Similar to Figures 2 and 3 except for the estimation was done based on daily and monthly slow investors only. ................................. 2

8 Model fit, semi-daily, daily, and monthly. Similar to Figures 2 and 3 except for the estimation was done based on semi-daily, daily, and monthly slow investors. ... 3

9 Model fit without leverage effect. Similar to Figures 2 and 3 except for removing the leverage effect channel (i.e., $\rho = 0$). ......................................................... 4
Figure 7. Model fit, daily and monthly. Similar to Figures 2 and 3 except for the estimation was done based on daily and monthly slow investors only.
Parameters best fit: $(\mu_{si}, \mu_{di}, \mu_{mi}) = (65.0, 1.0, 9.0), (\mu_{sr}, \mu_{dr}, \mu_{mr}) = (1.4, 0.0, 1.8), \beta_M = 0.015, \sigma_w = 4.8e-05, \sigma_w = 0.02, \rho = -0.18$

Figure 8. Model fit, semi-daily, daily, and monthly. Similar to Figures 2 and 3 except for the estimation was done based on semi-daily, daily, and monthly slow investors.
Parameters best fit: $(\mu_{si}, \mu_{di}, \mu_{mi}) = (65.0, 0.0, 9.0), (\mu_{sr}, \mu_{dr}, \mu_{mr}) = (1.4, 0.0, 2.0), \beta_M = 0.015, \beta_w = 0.00011, \sigma_w = 0.013, \rho = 0$

Figure 9. Model fit without leverage effect. Similar to Figures 2 and 3 except for removing the leverage effect channel (i.e., $\rho = 0$).