Estimating Information Asymmetry in Securities Markets

Kerry Back
Jones Graduate School of Business and Department of Economics
Rice University, Houston, TX 77005, U.S.A.

Kevin Crotty
Jones Graduate School of Business
Rice University, Houston, TX 77005, U.S.A.

Tao Li
Department of Economics and Finance
City University of Hong Kong, Kowloon, Hong Kong

Abstract

We propose and estimate a model of informed trading that is a hybrid of the PIN and Kyle models. The relationships between price impacts and the estimated parameters are consistent with the model’s theoretical predictions. We compare the hybrid model estimates to PIN estimates. The probability of information events in the PIN model is negatively correlated with the hybrid model’s estimated probability and with price impacts. A composite information asymmetry measure explains more cross-sectional variation in price impacts than does PIN, whose explanatory power stems primarily from its liquidity trading parameter. The empirical results are consistent with the model’s implication that both prices and order flows are necessary to identify private information.

\*Versions of this paper were presented under various titles at the University of Colorado, the SEC, the NYU Stern Microstructure Conference, and the University of Chicago Market Microstructure and High Frequency Data Conference. We thank Pete Kyle, Rob Engle, and seminar participants for helpful comments.

Email addresses: Kerry.E.Back@rice.edu (Kerry Back), Kevin.P.Crotty@rice.edu (Kevin Crotty), TaoLi3@cityu.edu.hk (Tao Li)

January 25, 2016
1. **Introduction**

Information asymmetry is a fundamental concept in economics, but its estimation is challenging because private information is generally unobservable. A large literature in finance and accounting utilizes the probability of informed trade (PIN) measure of Easley, Kiefer, O’Hara, and Paperman (1996) to proxy for information asymmetry because private information plays a key role in so many economic settings. However, a number of papers argue that PIN does not measure information asymmetry.\(^1\)

We propose a model of informed trading in securities markets that shares many features of the PIN model of Easley et al. (1996) but in which informed trading is endogenous as in Kyle (1985). We call this a hybrid PIN-Kyle model. We estimate the model and compare the parameter estimates to PIN estimates. We calculate a composite measure of information asymmetry, the expected average lambda (price impact), based on the underlying hybrid model parameters. This measure incorporates both the probability and magnitude of information events as well as the amount of liquidity trading.

The parameters of the hybrid model are estimated using both prices and order flows. On the other hand, the PIN parameters depend only on order flows. The hybrid model predicts that order flows alone cannot identify information asymmetry. The intuition is quite simple. Consider, for example, a stock for which there is a large amount of private information and another for which there is only a small amount of private information. If it is anticipated that private information is more of a concern for the first stock than for the second, then the first stock will be less liquid, other things being equal. The lower liquidity will reduce the amount of informed trading, possibly offsetting the increase in informed trading due to greater private information.

\(^1\)We review the literature in Section 2.
In equilibrium, the amount of informed trading may be the same in both stocks, despite the difference in information asymmetry. In general, the distribution of order flows need not reflect the degree of information asymmetry when liquidity providers react to information asymmetry and informed traders react to liquidity.

In the PIN model, order flows are, by assumption, independent of price changes. This is the reason price changes are not used in estimating the parameters. In the hybrid model, order flows depend on price changes, and both order flows and price changes are useful for estimating the parameters. Thus, the reaction of liquidity providers to information asymmetry and the reaction of informed traders to liquidity are taken into account when estimating the hybrid model.

Theory predicts that stock orders have larger price impacts when information events are more frequent or when information events are of larger magnitude. This is true for both the hybrid and PIN models. We estimate the price impacts of orders for a panel of stocks and regress price impacts on model parameters. Consistent with theory, price impacts are higher for stocks that have more frequent information events, larger magnitude events, or less volatile liquidity trading when these parameters are estimated with the hybrid model. Contrary to theory, price impacts are lower for stocks that have more frequent information events when the probability of information events is estimated with the PIN model. In fact, the estimates of the probability of an information event are negatively correlated across models.

Empirically, expected average lambda from the hybrid model explains a substantial amount of cross-sectional variation in price impacts. However, we also find that

---

2There seems to be general agreement that at least a portion of the price impact of trades is due to information asymmetry. Glosten and Harris (1988), Hasbrouck (1988), and Hasbrouck (1991) estimate models of trades and price changes in which both information asymmetry and inventory control motives are accommodated, and all three papers conclude that information asymmetry is important.
PIN is positively related to price impacts cross-sectionally, despite the negative correlation between the PIN probability of information events and price impacts. Economically, this is primarily due to the fact that PIN reflects its model’s estimate of the amount of liquidity trading. We run a horse race between the hybrid model’s expected average lambda and PIN by orthogonalizing PIN to expected average lambda. The component of PIN collinear with expected average lambda explains five times the variation in price impacts explained by the component of PIN orthogonal to expected average lambda.

2. Literature

We first describe some of the microstructure literature that is related to our theoretical model. We then review the literature that uses PIN as a proxy for information asymmetry as well as the literature critiquing the use of PIN as a measure of information asymmetry or as a statistical model of the order flow distribution.

2.1. Microstructure Theory

We analyze a model that includes some features of the PIN model but in which informed orders are endogenized as in Kyle (1985). Odders-White and Ready (2008) analyze a Kyle model in which the probability of an information event is less than 1, as it is in our model. However, Odders-White and Ready analyze a single-period model, whereas we analyze a dynamic model. In a single-period model, because of the net order having a mixture distribution, the conditional expectation of the asset value given the net order is not a linear function of the net order. To make the model tractable, Odders-White and Ready deviate from the usual Kyle model formulation and do not require the asset price to equal its conditional expected value. Instead, they only require that unconditional expected market maker profits are zero. They find the pricing rule that is linear in the net order that has this “zero conditional
expected profits on average” property. Such a pricing rule would require commitment by market makers, because it is not consistent with ex-post optimization by market makers. In contrast, pricing in our model is consistent with ex-post optimization by market makers: prices equal conditional expected values.

Related theoretical work includes Rossi and Tinn (2010), Banerjee and Green (2015), Foster and Viswanathan (1995), and Chakraborty and Yilmaz (2004). Rossi and Tinn solve a two-period Kyle model in which there are two large traders, one of whom is certainly informed and one of whom may or may not be informed. In their model, unlike ours, there are always information events. Banerjee and Green solve a rational expectations model with myopic mean-variance investors in which investors learn whether other investors are informed. They show that variation over time in the perceived likelihood of informed trading induces volatility clustering. While their model is quite different from ours, our model also exhibits volatility clustering. Volatility follows the same pattern as Kyle’s lambda, which varies over time due to variation in the market’s estimate of whether an information event occurred.

Foster and Viswanathan (1995) consider a series of single-period Kyle models in which traders choose in each period whether to pay a fee to become informed. There may be periods in which there are no informed traders. However, in their model, it is always common knowledge how many traders choose to become informed, so, in contrast to our model, there is no learning from orders about whether informed traders are present.

Chakraborty and Yilmaz (2004) study a discrete-time Kyle model in which there may or may not be an information event. Their main result is that the informed trader will manipulate (sometimes buying when she has bad information and/or selling when she has good information) if the horizon is sufficiently long. The primary difference between their model and ours is that in their model the noise trade distribution has
finite support. If the low type trader never buys in their model, then an aggregate order larger than the maximum of the noise trade distribution implies for certain that the low type trader is not present. When the horizon is sufficiently long, it is optimal for the low type trader to deviate from a strategy of never buying and to buy until the aggregate order in a period is large enough that market makers put 100% probability on her not being a low type. Then, she can begin selling. Consequently, it cannot be an equilibrium in their model for low types to never buy and for high types to never sell, when the horizon is sufficiently long. In contrast, market makers in our model can never rule out any type of the informed trader until the end of the model, so it does not strictly pay for a low type to pretend to be a high type or vice versa (as we will show, and as is true in other continuous-time Kyle models, the informed trader in our model is locally indifferent about buying or selling, so pretending to be a different type is not suboptimal, but it does not occur in equilibrium).

A precursor to our paper is Li (2012), which solves a continuous-time Kyle model in which the probability of an information event is less than 1 by applying filtering theory to a transformation of the aggregate order process. The filtering solution produces a stochastic differential equation for the equilibrium rather than a closed form solution. The method of proof used in this paper shares some features with the proof in Back and Crotty (2015).

2.2. PIN

A large literature in finance and accounting applies the PIN model to measure information asymmetry. A portion of those papers assesses whether information risk is priced. See, for example, Easley and O’Hara (2004), Duarte and Young (2009), Mohanram and Rajgopal (2009), Easley, Hvidkjaer, and O’Hara (2002), Easley, Hvidkjaer, and O’Hara (2010), Akins, Ng, and Verdi (2012), Li, Wang, Wu, and He (2009), and Hwang, Lee, Lim, and Park (2013). Many other papers use PIN as a proxy for
information asymmetry in a variety of applications ranging from corporate finance (e.g., Chen, Goldstein, and Jiang, 2007; Ferreira and Laux, 2007) to accounting (e.g., Frankel and Li, 2004; Jayaraman, 2008).

Critiques of PIN can be classified into two groups. One set of papers argues that it does not measure information asymmetry. The other set argues that its predictions for the distributions of order flows are inconsistent with empirical distributions. The first set of papers includes Aktas et al. (2007), Akay et al. (2012), and Duarte, Hu, and Young (2014). Aktas et al. (2007) examine trading around merger announcements. They show that PIN decreases prior to announcements. In contrast, percentage spreads and the permanent price impact of trades, measured as in Hasbrouck (1991), rise before announcements, indicating the presence of information asymmetry. They describe the decline in PIN prior to announcements as a PIN anomaly. Akay et al. (2012) show that PIN is higher in the Treasury bill market than it is in markets for individual stocks. Given that it is very doubtful that informed trading in T-bills is a frequent occurrence, this is additional evidence that PIN is not measuring information asymmetry. Duarte, Hu, and Young (2014) also examine merger announcements. They estimate the parameters of the PIN model and then compute the conditional probability of an information event each day. They show that the conditional probability rises prior to merger announcements but stays elevated for up to 30 days following announcements. They show that the high post-announcement conditional probabilities are due to high turnover and argue that high turnover is misidentified as private information by the PIN model.

Kim and Stoll (2014) analyze trading around earnings announcements and consider subsamples based on the magnitude of the surprise in the announcement. They find that order imbalances do not correlate with surprises and conclude that they are not the result of informed trading. The evidence regarding order imbalances and
earnings surprises is additional evidence that measures based on order imbalances, like PIN, cannot be measuring information asymmetry. However, the conclusion that imbalances are not due to informed trading seems unwarranted. It is based on the premise that informed order imbalances should be larger when surprises are larger, which ignores equilibrium reactions. If there is a greater ex-ante likelihood of informed trading in cases in which surprises are larger, then there will be less liquidity when surprises are larger, other things being equal, so informed order imbalances should not necessarily be larger when surprises are larger. Related evidence is provided by Collin-Dufresne and Fos (2015), who document that both PIN and price impacts are lower when activist investors accumulate positions in target firms ahead of required regulatory 13-D filings. Collin-Dufresne and Fos (2012) show that the negative relationship between liquidity and informed trading is the equilibrium outcome in a Kyle model due to informed traders reacting to changes in liquidity.

Venter and de Jongh (2006), Duarte and Young (2009), Gan, Wei, and Johnstone (2014), and Duarte, Hu, and Young (2014) all show that the PIN model fails to fit the empirical joint distribution of buy and sell orders. The first two papers propose variations of the PIN model designed to provide a better fit to the data.

Easley, López de Prado, and O’Hara (2012) propose a variation of PIN called VPIN (Volume-synchronized PIN). They argue that signing individual transactions as buys or sells in modern high-frequency markets may be problematic. Instead of attempting to sign individual transactions, they assign a fraction of total volume during a time interval to buys and the remaining fraction to sells based on the standardized price change during the time interval. They calculate VPIN using the order imbalance implied by these estimates of buys and sells. Easley et al. (2011) claim that VPIN predicted the “flash crash” of May 6, 2010. This claim and some other claims regarding VPIN are challenged by Andersen and Bondarenko (2014b). See
also Easley et al. (2014) and Andersen and Bondarenko (2014a).

Easley, Engle, O’Hara, and Wu (2008) model daily buys and sells as being generated from the PIN model with time-varying parameters. The parameters are assumed to be driven by buys and sells, much like the volatility in a GARCH model is an unobservable variable that depends on returns and then determines the distribution of future returns. When they estimate the model, they find that the informed and uninformed arrival rates both increase with volume, whether orders are balanced or unbalanced.

These variations on PIN are outside the scope of our paper. In comparing our information asymmetry estimates to PIN, our focus is on the methodology used in many papers in accounting and finance that use PIN as a proxy for information asymmetry and estimate the parameters using daily buys and sells as in Easley et al. (1996).  

3. The Hybrid Model

The hybrid model includes two important features of PIN models—a probability less than 1 of an information event and a binary asset value conditional on an information event—and it also includes an optimizing (possibly) informed trader, as in the Kyle (1985) model. Denote the time horizon for trading by $[0, 1]$. Assume there is a single risk-neutral strategic trader. Assume this trader receives a signal $S \in \{L, H\}$ at time 0 with probability $\alpha$, where $L < 0 < H$. Let $p_L$ and $p_H = 1 - p_L$ denote the probabilities of low and high signals, respectively, conditional on an information event. With probability $1 - \alpha$, there is no information event, and the trader also knows when this happens. Let $\xi$ denote an indicator for whether an information event has occurred ($\xi = 1$ if yes and $\xi = 0$ if no). In addition to the private information, public information can also arrive during the course of trading, represented by a martingale
Whether there was an information event, and, if so, whether the signal was low or high becomes public information after the close of trading, producing an asset value of $V_1 + \xi S$.

In addition to the strategic trades, there are liquidity trades represented by a Brownian motion $Z$ with zero drift and instantaneous standard deviation $\sigma$. Let $X_t$ denote the number of shares held by the strategic trader at date $t$ (taking $X_0 = 0$ without loss of generality), and set $Y_t = X_t + Z_t$. The processes $Y$ and $V$ are observed by market makers. Denote the information of market makers at date $t$ by $F_t^{V,Y}$.

One requirement for equilibrium in this model is that the price equal the expected value of the asset conditional on the market makers’ information and given the trading strategy of the strategic trader:

$$P_t = \mathbb{E} \left[ V_1 + \xi S \mid F_t^{V,Y} \right] = V_t + \mathbb{E} \left[ \xi S \mid F_t^{V,Y} \right]. \quad (1)$$

We will show that there is an equilibrium in which $P_t = V_t + p(t,Y_t)$ for a function $p$. This means that the expected value of $\xi S$ conditional on market makers’ information depends only on cumulative orders $Y_t$ and not on the entire history of orders.

The other requirement for equilibrium is that the strategic trades are optimal.

Let $\theta_t$ denote the trading rate of the strategic trader (i.e., $dX_t = \theta_t \, dt$). The process $\theta$ has to be adapted to the information possessed by the strategic trader, which is $V$, $\xi S$, and the history of $Z$ (in equilibrium, the price reveals $Z$ to the informed trader).

The strategic trader chooses the rate to maximize

$$\mathbb{E} \int_0^1 \left[ V_1 + \xi S - P_t \right] \theta_t \, dt = \mathbb{E} \int_0^1 \left[ \xi S - p(t,Y_t) \right] \theta_t \, dt, \quad (2)$$

with the function $p$ being regarded by the informed trader as exogenous. In the optimization, we assume that the strategic trader is constrained to satisfy the “no doubling strategies” condition introduced in Back (1992), meaning that the strategy
must be such that
\[ E \int_0^1 p(t, Y_t)^2 \, dt < \infty \]
with probability 1.

Let \( N \) denote the standard normal distribution function, and let \( n \) denote the standard normal density function. Set \( y_L = \sigma N^{-1}(\alpha p_L) \) and \( y_H = \sigma N^{-1}(1 - \alpha p_H) \). This means that the probability mass in the lower tail \((-\infty, y_L)\) of the distribution of cumulative liquidity trades \( Z_1 \) equals \( \alpha p_L \), which is the unconditional probability of bad news. Likewise, the probability mass in the upper tail \( (y_H, \infty) \) of the distribution of \( Z_1 \) equals \( \alpha p_H \), which is the unconditional probability of good news. Set

\[
q(t, y, s) = \begin{cases} 
E[Z_1 - Z_t \mid Z_t = y, Z_1 < y_L] & \text{if } s = L, \\
E[Z_1 - Z_t \mid Z_t = y, y_L \leq Z_1 \leq y_H] & \text{if } s = 0, \\
E[Z_1 - Z_t \mid Z_t = y, Z_1 > y_H] & \text{if } s = H.
\end{cases}
\]  

From the standard formula for the mean of a truncated normal, we obtain the following more explicit formula for \( q \):

\[
q(t, y, s) = \begin{cases} 
- n \left( \frac{y_L - y}{\sigma \sqrt{1-t}} \right) / N \left( \frac{y_L - y}{\sigma \sqrt{1-t}} \right) & \text{if } s = L, \\
\left[ n \left( \frac{y_L - y}{\sigma \sqrt{1-t}} \right) - n \left( \frac{y_H - y}{\sigma \sqrt{1-t}} \right) \right] / \left[ N \left( \frac{y_H - y}{\sigma \sqrt{1-t}} \right) - N \left( \frac{y_L - y}{\sigma \sqrt{1-t}} \right) \right] & \text{if } s = 0, \\
n \left( \frac{y - y_H}{\sigma \sqrt{1-t}} \right) / N \left( \frac{y - y_H}{\sigma \sqrt{1-t}} \right) & \text{if } s = H.
\end{cases}
\]  

The equilibrium described below can be shown to be the unique equilibrium in a certain broad class, following Back (1992).

**Theorem 1.** There is an equilibrium in which the trading rate of the strategic trader is

\[
\theta_t = \frac{q(t, Y_t, \xi S)}{1 - t}.
\]  

10
The equilibrium asset price is \( P_t = V_t + p(t, Y_t) \), where the pricing function \( p \) is given by

\[
p(t, y) = L \cdot N \left( \frac{y_L - y}{\sigma \sqrt{1 - t}} \right) + H \cdot N \left( \frac{y - y_H}{\sigma \sqrt{1 - t}} \right).
\]  

(6)

In this equilibrium, the process \( Y \) is a martingale given market makers’ information and has the same unconditional distribution as does the liquidity trade process \( Z \); that is, it is a Brownian motion with zero drift and standard deviation \( \sigma \).

The last statement of the theorem implies that the distribution of order flows in the model does not depend on the information asymmetry parameters \( \alpha \), \( H \), and \( L \). Thus, if the model is correct, it is impossible to estimate those parameters using order flows alone. In general, the theorem suggests that it may be difficult to identify information asymmetry parameters using order flows alone, as discussed in the introduction. When we estimate the hybrid model, we use both order flows and returns, in contrast to the PIN model that only uses order flows.

Empirically, we test the relationship between \( \alpha \) and price impacts of trades. Figure 1 plots the equilibrium price as a function of \( Y_t \) for two different values of \( \alpha \). It shows that the price is more sensitive to orders when \( \alpha \) is larger. This is also true in the PIN model. We test the relationship for both models. To investigate further how the sensitivity of prices to orders depends on \( \alpha \) in the hybrid model, we calculate the price sensitivity—that is, we calculate Kyle’s lambda.

**Theorem 2.** In the equilibrium of Theorem 1, the asset price evolves as \( dP_t = dV_t + \lambda(t, Y_t) \, dY_t \), where Kyle’s lambda is

\[
\lambda(t, y) = -\frac{L}{\sigma \sqrt{1 - t}} \cdot n \left( \frac{y_L - y}{\sigma \sqrt{1 - t}} \right) + \frac{H}{\sigma \sqrt{1 - t}} \cdot n \left( \frac{y - y_H}{\sigma \sqrt{1 - t}} \right)
\]

(7)

Furthermore, Kyle’s lambda \( \lambda(t, Y_t) \) is a martingale with respect to market makers’
information on the time interval $[0, 1)$.

Kyle’s lambda is a stochastic process in our model, but we can easily relate the expected average lambda to $\alpha$. Because lambda is a martingale, the expected average lambda is $\lambda(0, 0)$. Substitute the definitions of $y_L$ and $y_H$ in (7) to compute

$$
\lambda(0, 0) = -\frac{L}{\sigma} n \left( N^{-1}(\alpha p_L) \right) + \frac{H}{\sigma} n \left( N^{-1}(1 - \alpha p_H) \right).
$$

(8)

Figure 2 plots the expected average lambda as a function of $\alpha$ for two values of $H$, taking $L = -H$. Doubling the signal magnitudes doubles lambda. Furthermore, the expected average lambda is increasing in $\alpha$.

4. Parameter Estimates

We estimate the hybrid model as well as the PIN model. We use trade and quote data from TAQ for NYSE firms from 1993 through 2012. We sign trades as buys and sells using the Lee and Ready (1991) algorithm: trades above (below) the prevailing quote midpoint are considered buys (sells). If a trade occurs at the midpoint, then the trade is classified as a buy (sell) if the trade price is greater (less) than the previous differing transaction price. We use daily buys and sells to estimate the PIN model. For the hybrid model, we sample prices and order imbalances hourly and at the close and define order imbalances as shares bought less shares sold.

We estimate each of the models by maximum likelihood. For the PIN model, we

\[3\text{If information events occur for sure } (\alpha = 1), \text{ then } \lambda(0, 0) = (H - L) n(0)/\sigma. \text{ This is analogous to the result of Kyle (1985) that lambda is the ratio of the signal standard deviation to the standard deviation of liquidity trading. Of course, it is not quite the same as Kyle’s formula, because we have a binary signal distribution, whereas the distribution is normal in Kyle (1985).}

\[4\text{We require that firms have intraday trading observations for at least 200 days within the year. We also require firms have the same ticker throughout the year and experience no stock splits.}

\[5\text{Prior to 2000, quotes are lagged five seconds when matched to trades. For 2000-2002, quotes are lagged one second. From 2003 on, quotes are matched to trades in the same second. To account for quote updates within a second, we use the interpolated time technique introduced by Holden and Jacobsen (2014).} \]
assume each day is a separate realization of the model, as is standard. We estimate the models for each stock-year, assuming constant parameters within each year for each stock, as is also standard. The likelihood function is provided in Appendix B, as is the definition of PIN.

For the hybrid model, we also assume each day is a separate realization of the model, and we estimate the model for each stock-year, assuming constant parameters within each year for each stock. In each year for each stock, we assume there is a parameter $\kappa$ such that the possible signals on each trading day $i$ of the year are $H = -L = \kappa$. We assume that the possible signals are proportional to the observed opening price on day $i$, $P_{i0}$, so $\kappa$ is the magnitude of private information expressed as a return. We also assume the public information process $V$ is a geometric Brownian motion on each day with a constant volatility $\delta$. Appendix B derives the likelihood function for the hybrid model under these assumptions. The likelihood function depends on the signal magnitude $\kappa$, the probability $\alpha$ of information events, the probability $p_L$ of a negative signal conditional on an information event, the standard deviation $\sigma$ of liquidity trading, and the volatility $\delta$ of public information.

4.1. Estimates of the Hybrid Model

Figure 3 displays the time-series of average estimates and the interquartile range for the cross-section of stocks for the hybrid model. The average $\alpha$ is almost 70% in the early part of the sample and falls to about 50% by the end of the sample, indicating the likelihood of private information events, at least at the daily frequency we study, has fallen on average. This effect starts in 2007 and is evident in the decrease in the lower quartile of $\alpha$ estimates. The other components of private information events are the magnitude $\kappa$ of the signal and the likelihood $p_L$ of a bad event. Private information $\kappa$ initially rises during the late 1990s, but exhibits a strong downward trend thereafter. The average $p_L$ indicates that the distribution of information is
relatively symmetric between positive and negative events. We combine these estimates into a single composite measure of information asymmetry by calculating the expected average lambda from Equation 8. The estimates indicate that the amount of private information has fallen across the twenty year sample with the exception of the late 1990s and the financial crisis.

In general, the standard deviation $\sigma$ of order imbalances and the volatility $\delta$ of public information appear to be roughly stationary. Despite the well-documented rise of high-frequency trading and the associated sharp increase in trading volume, the estimated volatility of order imbalances has remained fairly stable over the twenty year sample. Public information volatility also spiked during the financial crisis. This suggests private information is generally proportional to public information rather than a fixed amount.

4.2. Testing Whether There is Always an Information Event in the Hybrid Model

Our hybrid model relaxes the assumption in Kyle (1985) of an information event each period. A natural question is whether this relaxation is supported in the data. The Kyle framework is nested in our model by the restriction that $\alpha = 1$. Accordingly, we estimate the model with this restriction. The standard likelihood ratio test of the null that $\alpha = 1$ against the alternative that $\alpha \in [0,1]$ is rejected for 75% of the firm-years (with a test size of 10%). However, the usual regularity conditions for the likelihood ratio test require that the restriction not be at the boundary of the parameter space. To address this issue, we bootstrap the distribution of the likelihood ratio statistic for a random sample of 100 firm-years as in Duarte and Young (2009).

Specifically, for a given firm-year, we estimate the restricted model ($\alpha = 1$) and then simulate 500 firm-years under the null using the estimated (restricted) parameters. We then estimate the restricted and unrestricted models for each simulated firm-year to obtain the distribution of the likelihood ratio under the null. The 90th
percentile of this distribution is the critical value to evaluate the empirical likelihood ratio. These bootstrapped likelihood ratio tests reject the restricted Kyle model in favor of the hybrid model for 76 of the 100 randomly selected firm-years. The data thus supports the conclusion that the probability of an information event is less than 1.

4.3. Estimates of the PIN Model

Figure 4 displays the time series of average parameter estimates for the PIN model. The average estimated $\alpha$ is much lower than in the hybrid model at 30 to 40%. The uninformed trading intensity $\varepsilon$ and informed trading intensity $\mu$ each rise markedly in the mid-2000’s reflecting the dramatic rise in trading volume. The average estimated PIN falls from about 15% in 1993 to 10% in 2012.

4.4. Comparing the Models

Table 1 reports correlations of the hybrid model parameters with those of the PIN model. The estimates of $\alpha$ for the hybrid model are negatively correlated with the PIN estimates of $\alpha$. The hybrid model $\alpha$ estimates are negatively correlated with the noise trading parameters across both models ($\sigma, \varepsilon$), and are also negatively correlated with the intensity of informed trading, $\mu$, in the PIN model. The hybrid $\alpha$ is positively correlated with PIN through its positive correlation with the ratio of informed to uninformed trading ($\frac{\mu}{\varepsilon}$).

The composite measures of information asymmetry for the models are expected average lambda and PIN. These measures are positively correlated (35%). This is surprising given the lack of positive correlations of the $\alpha$ measures.

A potential explanation lies in the modeling of noise trading. Equation 8 shows that expected average lambda is inversely related to the volatility of noise trading. The PIN measure is also inversely related to noise trading (Equation B.5). Empirically, the hybrid model order imbalance volatility, $\sigma$, is positively correlated with the
PIN noise trading intensity, $\varepsilon$, as expected but is also positively correlated with both the intensity of informed trading, $\mu$, and the probability of an information event estimated by the PIN model.\(^6\) One potential explanation for this is that the unconditional variance of order imbalances in the PIN model is:

$$2\varepsilon + \alpha \mu (1 + \mu) - (\alpha \mu (1 - 2p_L))^2$$

which is generally increasing in both $\alpha$ and $\mu$ for the values estimated. Since PIN is increasing in $\frac{\mu}{\varepsilon}$, the positive correlations of both the informed and uninformed trading intensities with $\sigma$ imply the relationship between PIN and $\sigma$ is unclear. Empirically, PIN and $\sigma$ are negatively correlated since $\frac{\mu}{\varepsilon}$ and $\sigma$ are negatively correlated. This negative correlation explains why PIN and expected average lambda are positively correlated despite very different estimates of the probability of an information event.

It is also interesting to see how estimates of private information differ in the cross-section of firms across models. Table 2 reports average values of the parameters within market capitalization deciles. For the hybrid model, the probability $\alpha$ of an information event, the magnitude $\kappa$ of information events, and expected average lambda are all larger for smaller firms. Thus, the hybrid model estimates indicate there is more private information for smaller firms. There is also more variation in the public valuation of smaller firms, as shown in the $\delta$ estimates.

In contrast, the estimates of the PIN model show smaller probabilities of information events for smaller firms. However, PIN is larger for smaller firms as a result of large increases in noise trading intensity $\varepsilon$ estimates as size increases. While the informed trading intensity $\mu$ also increases with size, the noise trading component dominates, leading to a declining ratio of informed to uninformed trading intensities.

\(^6\) $\mu$ and $\varepsilon$ themselves are very highly correlated.
as a function of size. This effect dominates the modest increases in $\alpha$ as a function of size, so PIN is lower as a result of increased noise trading identified by the PIN model.

5. Predicting Price Impacts

In theory, price impacts should be larger when information asymmetry is higher, for instance, when the probability $\alpha$ of information events is larger. This is true for both the hybrid and the PIN model. It is shown in Section 3 for the hybrid model. For the PIN model, the opening quoted spread is the product of PIN and the magnitude of the information, $H - L$.\(^7\) Note that PIN is increasing in $\alpha$ and $\mu$, and decreasing in $\varepsilon$. In this section, we assess how cross-sectional variation in price impacts relates to the estimated structural parameters from each model and functions of parameters theoretically related to spreads in each model.

The percent price impact of a given trade $k$ is calculated as:

\[
\text{Percent Price Impact}_k = \frac{2D_k(P_{k+5} - P_k)}{P_k},
\]

where $P_k$ is the prevailing quote midpoint for trade $k$, $P_{k+5}$ is the quote midpoint five minutes after trade $k$, and $D_k$ equals 1 if trade $k$ is a buy and $-1$ if trade $k$ is a sell. Goyenko, Holden, and Trzcinka (2009) use this measure as one of their high-frequency liquidity benchmarks in a study assessing the quality of various liquidity measures based on daily data.\(^8\) For a given stock-day, the estimate of the percent price impact is the equal-weighted average price impact over all trades on that day. We average these daily price impact estimates for each stock-year.

\(^7\)See Equation 11 of Easley et al. (1996), which assumes $p_L = p_H$.

\(^8\)Holden and Jacobsen (2014) show that liquidity measures such as the percent price impact can be biased when constructed from monthly TAQ data, so we follow their suggested technique in processing the data.
Figure 5 plots the time-series of the cross-sectional average and interquartile range of the estimated price impacts. Over the twenty year sample, price impacts initially rose over the 1990s before falling dramatically following the turn of the century, with the brief exception of the financial crisis. Note that the time-series of the hybrid model expected average lambda and the magnitude of private information, $\kappa$, exhibit similar patterns (Figure 3).

Table 3 reports univariate Fama and MacBeth (1973) cross-sectional regressions of price impacts on structural parameters in the hybrid and PIN models. Before running the regressions, the structural parameters are winsorized at 1/99% and standardized to have unit standard deviation. Price impacts are positively related to each of the hybrid model parameters that measure private information (the probability $\alpha$ of an information event and the magnitude $\kappa$ of information events), and price impacts are negatively related to the volatility of noise trading, $\sigma$. The expected average lambda implied by the estimates explains over 30% of the cross-sectional variation in price impacts and is strongly positively related to empirical price impacts.

Price impacts are higher for stocks with lower PIN $\alpha$ estimates. This is inconsistent with PIN $\alpha$ measuring the probability of private information events. Price impacts are also negatively correlated with the estimated trading intensities $\mu$ and $\varepsilon$. Theory predicts $\mu$ to be positively correlated with price impacts and $\varepsilon$ to be negatively correlated. PIN is increasing in the ratio of these intensities, $\mu/\varepsilon$, which correlates positively with price impacts as predicted by theory. Table 1 indicates that $\mu/\varepsilon$ is negatively correlated with both $\varepsilon$ and $\mu$ as well as with $\sigma$ from the hybrid model. Thus, PIN estimates capture the inverse relationship between noise trading and spreads. As a result, price impacts increase with PIN cross-sectionally in spite of the negative relationship between $\alpha$ and price impacts.

Table 4 reports multivariate cross-sectional regressions controlling for a firm’s size
and price using the logarithm of market capitalization and the inverse stock price as of the beginning of the year. Both size and price are strongly related to price impacts (Breen, Hodrick, and Korajczyk, 2002); on average, these two control variables explain 65% of the cross-sectional variation in price impact. Both expected average lambda from the hybrid model and PIN remain significantly positively correlated with price impacts, but the inclusion of size and price controls cuts the economic significance of both information asymmetry measures. A one standard deviation change in expected average lambda is associated with 2.25 basis point increase in price impact; the univariate association was about twice as large. The reduction is much more dramatic for PIN. A one standard deviation in PIN is associated with less than a basis point increase in price impacts. The magnitude of univariate relationship is almost 6 times the size of the multivariate relationship.

As discussed previously, expected average lambda from the hybrid model is positively correlated with PIN. How much of the variation in price impacts explained by these measures captures common information? To determine this, we orthogonalize PIN to expected average lambda via regression. This decomposition consists of the portion of PIN collinear with expected average lambda (denoted \( PIN \parallel \lambda \)) and the orthogonalized portion (denoted \( PIN \bot \lambda \)). Each component is scaled to have unit standard deviation. Column (4) of Table 4 shows the relationship between price impact and the decomposed PIN. The portion of PIN collinear with expected average lambda is strongly positively related to the cross-section of price impacts. The common component across measures explains about five times more variation in price impacts than the orthogonal component.

The theory in Section 3 shows that price impacts should be related to the inverse of the standard deviation of liquidity trading. Accordingly, we orthogonalize PIN to this measure and regress price impacts on the portion of PIN explained and unex-
plained by the inverse of noise trading along with the expected average lambda (last column of Table 4). Each component is standardized to have unit standard deviation. Expected average lambda again explains more of the cross-section of price impacts. The coefficient on lambda is about four times as large as that on the portion of PIN orthogonal to noise trading.

6. Further Comparison of the Models

In this section, we explain how the hypotheses of our model relate to those of the PIN model. This facilitates the explanation in the next subsection of why the endogeneity of informed orders in our model prevents the information asymmetry parameters from being identified from order flows alone.

In the PIN model, there is a probability $\alpha$ of an information event. If there is bad news, then informed sell orders arrive to the market as a Poisson process with parameter $\mu$. If there is good news, then informed buy orders arrive to the market as a Poisson process with parameter $\mu$. Regardless of whether there is an information event, uninformed buy orders and uninformed sell orders each arrive as Poisson processes with parameter $\varepsilon$. All of the Poisson processes are independent. Throughout the time period of the model, market makers revise their beliefs about whether an information event occurred and about the nature of any information event based on the order flows. These revisions are in accordance with Bayes’ rule. For example, if there has been a relatively large number of buy orders, then market makers believe it likely that there has been good news. Market makers also revise their bid and ask prices throughout the time period of the model as their beliefs change. These are simple calculations presented in Equations (4) and (5) of Easley et al. (1996).

An important feature of the PIN model, which makes it easy to estimate but which seems quite unrealistic, is that the informed order flow does not respond to
price changes. For example, when there is good news, informed buy orders continue to arrive as a Poisson process with parameter $\mu$ throughout the model, regardless of the price changes that have occurred previously. This is in part a consequence of the binary signal assumption: the bid and ask are always strictly between the bad news value and the good news value, so the asset is always somewhat overpriced when there is bad news and somewhat underpriced when there is good news. However, as illustrated in Figure 6, when informed orders are endogenized, the arrival rate of informed orders depends on prior price changes even with a binary signal. When prices have moved in the direction of the news, informed orders slow down, and, when prices have moved in the opposite direction, informed orders speed up.

A second key feature of the PIN methodology is that, while prices can easily be computed, they are not used to estimate the parameters of the model. The likelihood function that is maximized is the density function of the sample of daily buys and sells.

The primary differences between the hypotheses of the PIN model and the hypotheses of our model are:

- In the PIN model, informed trades and liquidity trades are independently distributed, conditional on an information event. However, in our model, the informed trader reacts to past liquidity trades—for example, by buying less of an undervalued asset when liquidity traders happen to erode the mispricing through net purchases of the asset.

- In the PIN model, the distribution of informed trades conditional on an information event is independent of the probability of an information event (they are governed by separate parameters). In our model, if the probability of an information event is larger, then Kyle’s lambda is larger, and the informed trader
reacts to the decrease in liquidity by trading less.

- In the PIN model, liquidity-motivated buys and sells are independent Poisson processes. In our model, net liquidity trades are a Brownian motion.

- In the PIN model, when there is no information event, all of the trades are liquidity trades. In our model, when there is no information event, there is a trader who recognizes that fact and who trades as a contrarian, selling when liquidity traders buy and drive the price up and buying when they sell.

The changes to the PIN model described in the first two bullet points seem very reasonable. Informed traders should trade less if liquidity traders happen to erode the mispricing and should trade more when the opposite happens. Furthermore, they should trade less in illiquid markets and more in liquid markets.

The third bullet point seems to be a relatively unimportant difference between the models. If Poisson buys and sells are small and arrive frequently, then net orders (buys minus sells) will be approximately a Brownian motion. The chief difference between the Brownian and Poisson models is that buys and sells are not separately defined in the Brownian model (a Brownian motion has infinite total variation, so it is not the difference of two increasing processes). However, recent variations of PIN (see the discussion of VPIN in Section 2) use only order imbalances—rather than separate counts of buys and sells—to attempt to measure information asymmetry. Our negative result on identification is directly applicable to those versions of PIN. More generally, if information asymmetry cannot be identified from order imbalances, as is true in our model, then it seems likely that it could not be identified from separate counts of buys and sells either.

---

9Back and Baruch (2004) show that the equilibrium of a Kyle (1985) model with Brownian orders can be approximated by a Glosten and Milgrom (1985) model with Poisson orders.
The change described in the fourth bullet point may be more controversial. The existence of contrarian traders as described in the fourth bullet point seems likely if there are always some traders who are best informed—corporate managers, for example. This would be the case if information is truly idiosyncratic to the firm. If, on the other hand, there is an industry or other aggregate component to the information, then it is possible that no one knows when no one else has information. In that case, the contrarian trader we postulate would not exist. Thus, we do not claim that our model is applicable in all circumstances. Nevertheless, it appears to be reasonable in some circumstances, and the empirical analysis in Sections 4 and 5 suggest that the hybrid model’s structural estimates do capture information asymmetry.

6.1. Identification using Order Flows Alone

The key implication of Theorem 1 for the identification of information asymmetry from order flows is the fact that the aggregate order imbalance $Y_1$ is normally distributed with mean zero and variance equal to the variance of liquidity trades. Applications of the PIN model typically assume each day is a separate instance of the model and use daily buys and sells to estimate the model parameters. If our model describes reality and each day is a separate instance of the model, then the sample of daily order imbalances is a sample of i.i.d. normal random variables with mean zero and variance equal to the variance of liquidity trades.\textsuperscript{10} The distribution of the sample is invariant with respect to the frequency and magnitude of information events, so the sample of daily order imbalances cannot identify information asymmetry.

The fact that $Y_1$ is normally distributed with mean zero and variance equal to the variance of liquidity trades is a consequence of the martingale property of $Y$ (a

\textsuperscript{10}Of course, we cannot know if a single day is a separate instance of either model. Many (long-lived) instances of private information may entail informed trading over multiple days (e.g., activist investors in Collin-Dufresne and Fos (2015)).
continuous martingale with quadratic variation over each time interval equal to the length of the interval is automatically a Brownian motion). The martingale property of $Y$ is equivalent to unpredictability of informed orders in our model. As mentioned before, informed orders are predictable in the PIN model, because informed traders do not react to price changes in the PIN model.

The negative result on identification also holds in a more general model in which there is a predictable component of order flows. In that model,

$$Y_1 = \int_0^1 \mu_t \, dt + Y_1^*,$$

where $Y^*$ is the sum of informed orders and unpredictable liquidity trades, and where $\mu$ is adapted to the price process and hence adapted to $Y^*$. Because informed orders are unpredictable, $Y^*$ is a martingale; therefore, it is a Brownian motion with its variance determined by the variance of liquidity trades. This implies that the distribution of $Y_1$ is again invariant with respect to the information asymmetry parameters.

Further insight into the identification issue can be gained by noting that, as in the PIN model, the unconditional distribution of the order imbalance in our model is a mixture of three conditional distributions. With probability $\alpha_p L$, $Y_1$ is drawn from the distribution conditional on a low signal; with probability $\alpha_p H$, $Y_1$ is drawn from the distribution conditional on a high signal; and with probability $1 - \alpha$, $Y_1$ is drawn from the distribution conditional on no information event. The first two distributions have nonzero means—there is an excess of sells over buys in the first and an excess of buys over sells in the second. This is also analogous to the PIN model. Thus, one might conjecture that changing $\alpha$—thereby changing the likelihood of drawing from the first two distributions—will alter the unconditional distribution of $Y_1$. If so, then one could perhaps identify $\alpha$ from the distribution of $Y_1$. In the PIN model, it is indeed true that changing $\alpha$, holding other parameters constant, alters the unconditional
distribution of the order imbalance. However, it is not true in our model, because
the distribution of informed trades in our model depends endogenously on $\alpha$ due to
liquidity depending on $\alpha$. With a larger alpha, the market is less liquid (see Section 3)
and the informed trader trades less aggressively. Thus, the distributions over which we
are mixing change when the mixture probabilities change, leaving the unconditional
distribution of $Y_1$ invariant with respect to $\alpha$.

The change in the conditional distributions is illustrated in Figure 7. The top and
bottom panels of Figure 7 show that the strategic trader trades more aggressively
when an information event occurs if an information event is less likely ($\alpha = 0.1$ versus
$\alpha = 0.5$). This equilibrium reaction of informed trading to exogenous changes in the
probability of information events is missing in the PIN model, in which informed
trading is exogenously determined. It is a key feature of our model that results in
the probability of information events being unidentified by the distribution of order
imbalance. The unconditional distribution of $Y_1$ is standard normal for both $\alpha = 0.1$
and $\alpha = 0.5$ in Figure 7, so we cannot hope to use the unconditional distribution to
recover $\alpha$.

Continuing with the example in Figure 7, calculate the expected absolute order imbalance as

$$ \alpha p_L E[|Y_1| \mid \xi = 1, S = L] + (1 - \alpha) E[|Y_1| \mid \xi = 0] + \alpha p_H E[|Y_1| \mid \xi = 1, S = H], $$

with $\sigma = 1$ and $p_L = p_H = 1/2$. If $\alpha = 0.5$, then the expected absolute order imbalance is

$$ \frac{1}{4} \times 1.27 + \frac{1}{2} \times 0.32 + \frac{1}{4} \times 1.27 = 0.80. $$

On the other hand, if $\alpha = 0.1$, then the expected absolute order imbalance is

$$ \frac{1}{20} \times 2.06 + \frac{9}{10} \times 0.66 + \frac{1}{20} \times 2.06 = 0.80. $$
Again, we see that informed trading is more aggressive when information events occur if such events are less likely. Here, it is clear that the endogenous change in informed trading exactly offsets the exogenous change in the likelihood of information events. In other words, the endogenous changes in the distributions over which we are mixing exactly offset the changes in the mixing probabilities (this is true even with $p_L \neq p_H$).

On the other hand, under the assumptions of the PIN model, the expected absolute order imbalance varies with $\alpha$ (see Easley et al., 2008, p. 176, for a discussion of how the absolute order imbalance is related to $\alpha$ and to PIN under the assumptions of the PIN model).

The previous paragraphs describe the invariance of the unconditional distribution of $Y_1$ with respect to $\alpha$. The other important parameters governing information asymmetry are $L$ and $H$. For example, if the possible signals $L$ and $H$ are both small in absolute value, then information asymmetry is a minor concern even if information events occur frequently. Order flows cannot identify $L$ and $H$ in our model. In fact, $L$ and $H$ do not affect even the conditional distributions shown in Figure 7; thus, they do not affect the unconditional distribution of $Y_1$.

Of course, identifying the information asymmetry parameters from the distribution of order imbalances is a very different issue from using order imbalances to update the probability of an information event in a particular instance of the model. Conditional on knowledge of the parameters, the order imbalance does help in estimating whether an information event occurred in a particular instance of the model; in fact, the market makers in the model update their beliefs regarding the occurrence of an information event based on the order imbalance. So, we can compute

$$\text{prob}(\text{info event} | Y_1, \text{parameters})$$

and this probability does depend on the information asymmetry parameters. We
could use this to identify the information asymmetry parameters if we had data on order imbalances and data on whether information events occurred. Of course, we do not have data of the latter type. Theorem 1 shows that the likelihood function of the information asymmetry parameters given only data on order imbalances is a constant function of those parameters; hence, the order imbalances alone cannot identify them.

6.2. Empirical and Model-Implied Unconditional Order Flow Distributions

The hybrid and PIN models have different implications for the unconditional distribution of order imbalances. For both models, the order flow distribution is a mixture distribution (i.e., of the good, bad, no event distributions), but Figure 8 shows how the distributions can differ based on the underlying parameter values. Under the hybrid model, end-of-day order flows are normally distributed with standard deviation $\sigma$. However, in the PIN model, order imbalances can be trimodal. Indeed, this is generally the case for order imbalances implied by structural estimates of the PIN model. The bottom row of Figure 8 plots the model-implied order imbalance distribution for the smallest and largest NYSE firm deciles under PIN. The PIN model must fit volume as well as order imbalances since the input data are buy and sell volumes. On the other hand, the hybrid model need only fit the order flow distribution.

How do the model-implied order imbalance distributions compare to those found empirically? Figure 9 shows the empirical standardized order imbalance distributions for the smallest and largest NYSE size deciles in our sample. The figure displays both share and trade imbalances since these are the underlying data for the hybrid and PIN models, respectively. The empirical distributions do not exhibit strong multimodal behavior. This is more consistent with the modeling assumption of the hybrid model than that of the PIN model.
7. Conclusion

We propose a model of informed trading that is a hybrid of the PIN and Kyle models. Information events occur with probability less than one as in the PIN model (but unlike the Kyle model). However, unlike the PIN model, informed orders are endogenously determined as in the Kyle model. We estimate the model and compare the parameter estimates to PIN estimates. The relationships between price impacts and the estimated hybrid model parameters are consistent with the theoretical predictions of the model. This is not true for the PIN parameters. The probability of information events in the PIN model is negatively correlated with price impacts and negatively correlated with the probability of information events estimated from the hybrid model. A composite measure of the hybrid model information asymmetry parameters explains much more cross-sectional variation in price impacts than does PIN. We also show that the explanatory power of PIN for price impacts is due primarily to the liquidity trading parameter. The improved empirical performance likely results from the use of both price and order flow data in estimation. Indeed, the model implies that both prices and order flows are needed to identify private information.
Appendix A. Proofs

The process $Y$ described in the following lemma is a variation of a Brownian bridge. It differs from a Brownian bridge in that the endpoint is not uniquely determined but instead is determined only to lie in an interval—either the lower tail $(-\infty, y_L)$, the upper tail $(y_H, \infty)$ or the middle region $[y_L, y_H]$—depending on whether there is an information event and whether the news is good or bad. Part (C) of the lemma follows immediately from the preceding parts, because the probability (A.3) is the probability that $Y_1 \notin [y_L, y_H]$ calculated on the basis that $Y$ is an $\mathbb{F}^Y$–Brownian motion with zero drift and standard deviation $\sigma$.

**Lemma.** Let $N$ denote the standard normal distribution function. Let $\mathbb{F}^Y = \{\mathcal{F}^Y_t \mid 0 \leq t \leq 1\}$ denote the filtration generated by the stochastic process $Y$ defined by $Y_0 = 0$ and
\[
dY_t = \frac{q(t, Y_t, \xi S)}{1 - t} \, dt + dZ_t. \tag{A.1}
\]
Then, the following are true:

(A) $Y$ is an $\mathbb{F}^Y$–Brownian motion with zero drift and standard deviation $\sigma$.

(B) With probability one,
\[
\begin{align*}
\xi = 1 \text{ and } S = L & \Rightarrow Y_1 < y_L, \quad \tag{A.2a} \\
\xi = 0 & \Rightarrow y_L \leq Y_1 \leq y_H, \quad \tag{A.2b} \\
\xi = 1 \text{ and } S = H & \Rightarrow Y_1 > y_H. \quad \tag{A.2c}
\end{align*}
\]

(C) For each $t < 1$, the probability that $\xi = 1$ conditional on $\mathcal{F}^Y_t$ is
\[
N \left( \frac{y_L - Y_1}{\sigma \sqrt{1 - t}} \right) + 1 - N \left( \frac{y_H - Y_1}{\sigma \sqrt{1 - t}} \right). \tag{A.3}
\]
Proof of the Lemma. Set

\[ k(1, y, s) = \begin{cases} 
1_{\{y < y_L\}} & \text{if } s = L, \\
1_{\{y_L \leq y \leq y_H\}} \, dz & \text{if } s = 0, \\
1_{\{y > y_H\}} & \text{if } s = H, 
\end{cases} \]

and, for \( t < 1, \)

\[ k(t, y, s) = \begin{cases} 
N\left(\frac{y_L - y}{\sigma \sqrt{1-t}}\right) \, dz & \text{if } s = L, \\
N\left(\frac{y_L - y}{\sigma \sqrt{1-t}}\right) - N\left(\frac{y_H - y}{\sigma \sqrt{1-t}}\right) \, dz & \text{if } s = 0, \\
N\left(\frac{y - y_H}{\sigma \sqrt{1-t}}\right) & \text{if } s = H. 
\end{cases} \]

Define

\[ \ell(t, y, s) = \frac{\partial \log k(t, y, s)}{\partial y}, \]

for \( t < 1. \) Then, \((1 - t)\sigma^2 \ell(t, y, s) = q(t, y, s)\) for \( t < 1, \) and the stochastic differential equation (A.1) can be written as

\[ dY_t = \sigma^2 \ell(t, Y_t, \xi S) \, dt + dZ_t \quad (A.4) \]

The process \( Y \) is an example of a Doob \( h \)-transform—see Rogers and Williams (2000).

To put (A.4) in a more standard form, define the two-dimensional process \( \hat{Y}_t = (\xi S, Y_t) \) with random initial condition \( \hat{Y}_0 = (\xi S, 0), \) and augment (A.4) with the equation \( d(\xi S) = 0. \) The existence of a unique strong solution \( \hat{Y} \) to this enlarged system follows from Lipschitz and growth conditions satisfied by \( \ell. \) See Karatzas and Shreve (1988, Theorem 5.2.9).

The uniqueness in distribution of weak solutions of stochastic differential equations
(Karatzas and Shreve, 1988, Theorem 5.3.10) implies that we can demonstrate Properties (A) and (B) by exhibiting a weak solution for which they hold. To construct such a weak solution, define a new measure $Q$ on $\mathcal{F}_1$ using $k(1, Z_t, \xi S)/k(0,0,\xi S)$ as the Radon-Nikodym derivative. The definition of $k$ implies that $k(t, Z_t, \xi S)$ is the $\mathcal{F}_t$-conditional expectation of the indicator function $k(1, Z_1, \xi S)$, so $k(t, Z_t, \xi S)$ is a martingale on the filtration $\mathcal{F}$. By Girsanov’s Theorem, the process $Z^*$ defined by $Z^*_0 = 0$ and
\[
dZ^*_t = -\sigma^2 \ell(t, Z_t, \xi S) \, dt + dZ_t
\]
is a Brownian motion (with zero drift and standard deviation $\sigma$) on the filtration $\mathcal{F}$ relative to $Q$. It follows that $Z$ is a weak solution of (A.4) relative to the Brownian motion $Z^*$ on the filtered probability space $(\Omega, \mathcal{F}, Q)$.

To establish Property (A) for the weak solution, we need to show that $Z$ is a Brownian motion on $(\Omega, \mathcal{G}, Q)$. Because $Z$ is a Brownian motion on $(\Omega, \mathcal{G}, P)$, it suffices to show that $Q = P$ when both are restricted to $\mathcal{G}_1$. This holds if for all $t_1 < \cdots < t_n \leq 1$ and all Borel $B$ we have
\[
P((Z_{t_1}, \ldots, Z_{t_n}) \in B) = Q((Z_{t_1}, \ldots, Z_{t_n}) \in B). \tag{A.5}
\]
The right-hand side of (A.5) equals
\[
E \left[ \frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} 1_B(Z_{t_1}, \ldots, Z_{t_n}) \right],
\]
which can be represented as the following sum:

\[
\alpha_p E \left[ \frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} 1_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi S = L \right] \\
+ (1 - \alpha) E \left[ \frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} 1_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi = 0 \right] \\
+ \alpha_p H \left[ \frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} 1_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi S = H \right].
\]

Using the definitions of \( y_L, y_H, \) and \( k, \) this equals

\[
E \left[ 1_{\{Z_1 < y_L\}} 1_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi S = L \right] \\
+ E \left[ 1_{\{y_L \leq Z_1 \leq y_H\}} 1_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi = 0 \right] \\
+ E \left[ 1_{\{Z_1 > y_H\}} 1_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi S = H \right].
\]

The \( \mathbb{P} \)-independence of \( Z \) and \( \xi S \) imply that the conditional expectations equal the unconditional expectations, so adding the three terms gives

\[
E[1_B(Z_{t_1}, \ldots, Z_{t_n})] = \mathbb{P}((Z_{t_1}, \ldots, Z_{t_n}) \in B).
\]

This completes the proof that \( Z \) is a Brownian motion on \((\Omega, \mathcal{G}, \mathbb{Q})\).

To establish Property (B) for the weak solution of \((A.4)\), we need to show that

\[
\mathbb{Q}(Z_1 < y_L \mid \xi S = L) = 1,
\]

\[
\mathbb{Q}(y_L \leq Z_1 \leq y_H \mid \xi = 0) = 1,
\]

\[
\mathbb{Q}(Z_1 > y_H \mid \xi S = H) = 1.
\]
Consider (A.6a). We have

\[
Q(\xi_S = L) = E \left[ \frac{k(1, Z_1, \xi_S)}{k(0, 0, \xi_S)} 1_{\{\xi_s = L\}} \right] \\
= E \left[ \frac{k(1, Z_1, L)}{k(0, 0, L)} 1_{\{\xi_s = L\}} \right] \\
= E \left[ 1_{\{Z_1 < y_L\}} 1_{\{\xi_s = L\}} \right] / \alpha p_L \\
= \alpha p_L,
\]

using the definition of \(k\) for the third equality and the \(\mathbb{P}\)-independence of \(Z\) and \(\xi_S\) for the last equality. By similar reasoning,

\[
Q(Z_1 < y_L, \xi_S = L) = E \left[ \frac{k(1, Z_1, \xi_S)}{k(0, 0, \xi_S)} 1_{\{Z_1 < y_L\}} 1_{\{\xi_s = L\}} \right] \\
= E \left[ \frac{k(1, Z_1, L)}{k(0, 0, L)} 1_{\{Z_1 < y_L\}} 1_{\{\xi_s = L\}} \right] \\
= E \left[ 1_{\{Z_1 < y_L\}} 1_{\{\xi_s = L\}} \right] / \alpha p_L \\
= \alpha p_L.
\]

Thus,

\[
Q(Z_1 < y_L \mid \xi_S = L) = \frac{Q(Z_1 < y_L, \xi_S = L)}{Q(\xi_S = L)} = \frac{\alpha p_L}{\alpha p_L} = 1.
\]

Conditions (A.6b) and (A.6c) can be verified by the same logic.

Proof of Theorem 1. It is explained in the text why the equilibrium condition (1) holds. It remains to show that the strategy (5) is optimal for the informed trader.

Let \(G \overset{\text{def}}{=} \{G_t \mid 0 \leq t \leq T\}\) denote the completion of the filtration generated by \(Z\), form the enlarged filtration with \(\sigma\)-fields \(G_t \vee \sigma(\xi_S)\), and let \(F \overset{\text{def}}{=} \{F_t \mid 0 \leq t \leq T\}\) denote the completion of the enlarged filtration. The filtration \(F\) represents the informed trader’s information.

33
Define

\[
J(1, y, L) = -L(y - y_L)1_{\{y > y_L\}} + H(y - y_H)1_{\{y > y_H\}},
\]
\[
J(1, y, 0) = -L(y_L - y)1_{\{y < y_L\}} + H(y - y_H)1_{\{y > y_H\}},
\]
\[
J(1, y, H) = -L(y_L - y)1_{\{y < y_L\}} + H(y_H - y)1_{\{y < y_H\}}.
\]

For \( t < 1 \) and \( s \in \{L, 0, H\} \), set \( J(t, y, s) = \mathbb{E}[J(t, Z_1, s) \mid Z_t = y] \). Then, \( J(t, Z_t, \xi S) \) is an \( \mathbb{F} \)-martingale, so it has zero drift. From Itô’s formula, its drift is

\[
\frac{\partial}{\partial t}J(t, Z_t, \xi S) + \frac{1}{2} \sigma^2 \frac{\partial^2}{\partial z^2}J(t, Z_t, \xi S).
\]

Equating this to zero, Itô’s formula implies

\[
J(1, Y_1, \xi S) = J(0, 0, \xi S) + \int_0^1 dJ(t, Y_t, \xi S) = J(0, 0, \xi S) + \int_0^1 \frac{\partial J(t, Y_t, \xi S)}{\partial y} dY_t.
\]

Therefore,

\[
\mathbb{E}[J(1, Y_1, \xi S) - J(0, 0, \xi S)] = \mathbb{E} \int_0^1 \frac{\partial J(t, Y_t, \xi S)}{\partial y} dY_t.
\]

(A.7)

To calculate \( \frac{\partial J(t, y, s)}{\partial y} \), use the fact that, by independent increments,

\[
J(t, y, s) = \mathbb{E}[J(t, Z_1, s) \mid Z_t = y] = \mathbb{E}[J(t, Z_1 - Z_t + y, s)]
\]

to obtain

\[
\frac{\partial J(t, y, s)}{\partial y} = \mathbb{E} \left[ \frac{\partial}{\partial y} J(t, Z_1 - Z_t + y, s) \right].
\]
Now, note that, for any real number \( a \) excluding the kinks at \( y_L - y \) and \( y_H - y \),

\[
\frac{\partial}{\partial y} J(1, a + y, L) = -L \mathbb{1}_{\{a > y_L - y\}} + H \mathbb{1}_{\{a > y_H - y\}} ,
\]

\[
\frac{\partial}{\partial y} J(1, a + y, 0) = L \mathbb{1}_{\{a < y_L - y\}} + H \mathbb{1}_{\{a > y_H - y\}} ,
\]

\[
\frac{\partial}{\partial y} J(1, a + y, H) = L \mathbb{1}_{\{a < y_L - y\}} - H \mathbb{1}_{\{a < y_H - y\}} .
\]

Therefore,

\[
\frac{\partial J(t, y, L)}{\partial y} = - L N \left( \frac{y - y_L}{\sigma \sqrt{1 - t}} \right) + H N \left( \frac{y - y_H}{\sigma \sqrt{1 - t}} \right) ,
\]

\[
\frac{\partial J(t, y, 0)}{\partial y} = L N \left( \frac{y - y_L}{\sigma \sqrt{1 - t}} \right) + H N \left( \frac{y - y_H}{\sigma \sqrt{1 - t}} \right) ,
\]

\[
\frac{\partial J(t, y, H)}{\partial y} = L N \left( \frac{y - y_L}{\sigma \sqrt{1 - t}} \right) - H N \left( \frac{y - y_H}{\sigma \sqrt{1 - t}} \right) .
\]

Now, the definition (6) gives us

\[
\frac{\partial J(t, y, s)}{\partial y} = \pi(t, y) - s
\]

for all \( s \in \{L, 0, H\} \). Substituting this into (A.7) gives us

\[
\mathbb{E}[J(1, Y_1, \xi S) - J(0, 0, \xi S)] = \mathbb{E} \int_0^1 [\pi(t, Y_t) - \xi S] dY_t .
\]

(A.8)

The “no doubling strategies” condition implies that \( \int p \, dZ \) is a martingale, so the right-hand side of this equals

\[
\mathbb{E} \int_0^1 [\pi(t, Y_t) - \xi S] \theta_t \, dt .
\]
Rearranging produces
\[ E \int_0^1 [\xi S - p(t, Y_t)] \theta_t \, dt = E[J(0, 0, \xi S) - J(1, Y_1, \xi S)] \leq E[J(0, 0, \xi S)], \]
using the fact that \( J(1, y, s) \geq 0 \) for all \((y, s)\) for the inequality. Thus, \( E[J(0, 0, \xi S)] \) is an upper bound on the expected profit, and the bound is achieved if and only if \( J(1, Y_1, \xi S) = 0 \) with probability one. By the definition of \( J(1, y, s) \), this is equivalent to \( Y_1 < y_L \) with probability one when \( \xi S = L, y_L \leq Y_1 \leq y_H \) with probability one when \( \xi = 0 \), and \( Y_1 > y_H \) with probability one when \( \xi S = H \). By part (B) of the proposition, the strategy (5) is therefore optimal.

\[ \square \]

Proof of Theorem 2. By Itô’s formula and the fact that \((dY)^2 = (dZ)^2 = \sigma^2 \, dt\), we have
\[ dp(t, Y_t) = \left( p_t(t, Y_t) + \frac{1}{2} \sigma^2 \partial_{yy} p(t, Y_t) \right) \, dt + p_y(t, Y_t) \, dY_t, \]
where we use subscripts to denote partial derivatives. Both \( Y \) and \( p(t, Y_t) \) are martingales with respect to the market makers’ information, so the drift term must be zero. That can also be verified by direct calculation of the partial derivatives, using the formula (6) for \( p(t, y) \). Thus,
\[ dp(t, Y_t) = p_y(t, Y_t) \, dY_t. \]

A direct calculation based on the formula (6) for \( p(t, y) \) shows that \( p_y(t, y) = \lambda(t, y) \) defined in (7).

To see that \( \lambda(t, Y_t) \) is a martingale for \( t \in [0, 1] \), with respect to market makers’
information, we can calculate, for $t < u < 1$,

$$
E[\lambda(u, Y_u) \mid Y_t = y] = -\frac{L}{\sigma \sqrt{1 - u}} \cdot \int_{-\infty}^{\infty} n\left( \frac{y_L - y'}{\sigma \sqrt{1 - u}} \right) f(y' \mid u - t, y) dy' 
+ \frac{H}{\sigma \sqrt{1 - u}} \cdot \int_{-\infty}^{\infty} n\left( \frac{y_H - y'}{\sigma \sqrt{1 - u}} \right) f(y' \mid u - t, y) dy',
$$

where $f(\cdot \mid \tau, y)$ denotes the normal density function with mean $y$ and variance $\sigma^2 \tau$.

A straightforward calculation shows that this equals $\lambda(t, y)$. For example, to evaluate the first term, use the fact that

$$
\frac{1}{\sigma \sqrt{1 - u}} n\left( \frac{y_L - y'}{\sigma \sqrt{1 - u}} \right) f(y' \mid u - t, y) 
= \frac{1}{\sigma \sqrt{1 - t}} n\left( \frac{y_L - y}{\sigma \sqrt{1 - t}} \right) \times \frac{1}{\sqrt{2\pi \sigma^2 (1 - u)(u - t)/(1 - t)}} 
\times \exp \left( - \frac{1 - t}{2(1 - u)(u - t)\sigma^2} \left( y' - \frac{(1 - u)y + (u - t)y_L}{1 - t} \right)^2 \right),
$$

which integrates to

$$
\frac{1}{\sigma \sqrt{1 - t}} n\left( \frac{y_L - y}{\sigma \sqrt{1 - t}} \right),
$$

because the other factors constitute a normal density function. \hfill \square
Appendix B. Likelihood Functions

Appendix B.1. Hybrid Model

Assume the trading period [0, 1] corresponds to a day. This implies that any private information becomes public before trading opens on the following day. We can estimate the model parameters using intraday price and order flow information. If we assume further that the model parameters are stable over time, then the price and order flow information from multiple days can be merged to estimate the parameters with greater precision.

The opening price on each day $i$ is $P_{i0} \overset{\text{def}}{=} \mathbb{E}[V_i + \xi_i S_i] = V_{i0}$. To obtain stationarity, we assume that the signal $S_i$ on day $i$ is proportional to the observed opening price $P_{i0}$. This construction causes the pricing function to be day-specific, and we denote it by $p_i(t, y)$. In fact,

$$p_i(t, y) = P_{i0} \times p(t, y)$$

where $p(t, y)$ is defined in Theorem 1. We specify that $H = -L = \kappa$ in the empirical implementation. Under this specification, the pricing function expressed in returns is given by:

$$p(t, Y_t) = \begin{cases} 
-\kappa 1(Y_t < z_L) + \kappa 1(Y_t > z_H) & \text{if } t = 1, \\
-\kappa F(z_L| t, Y_t) + \kappa (1 - F(z_H| t, Y_t)) & \text{if } t < 1 
\end{cases}$$

where $F(z| t, Y_t)$ is the normal distribution function with mean $Y_t$ and variance $(1 - t)\sigma^2$.

\footnote{In contrast to Odders-White and Ready (2008), our estimation does not use overnight returns. In our theoretical model, private information that is made public after the close of trading is incorporated into prices before trading ends (convergence to strong-form efficiency). Thus, overnight returns in our model are due to arrival of new public information, which does not aid in estimating the model.}
The price at time $t$ on day $i$ is $V_{it} + p_i(t, Y_{it})$, so the gross return through time $t$ is

$$\frac{P_{it}}{P_{i0}} = \frac{V_{it}}{V_{i0}} + \frac{p_i(t, Y_{it})}{P_{i0}} = \frac{V_{it}}{V_{i0}} + p(t, Y_{it}). \quad (B.1)$$

Assume

$$\frac{dV_{it}}{V_{it}} = \delta \, dB_{it}$$

for a constant $\delta$ and a Brownian motion $B_i$, so we have

$$\frac{P_{it}}{P_{i0}} = p(t, Y_{it}) + e^{\delta B_{it} - \delta^2 t/2}.$$ 

Assume the price and order imbalance are observed at times $t_1, \ldots, t_{k+1}$ each day with $t_{k+1} = 1$ being the close and the other times being equally spaced: $t_j = j \Delta$ for $\Delta > 0$ and $j \leq k$. Let $P_{ij}$ denote the observed price and $Y_{ij}$ the observed order imbalance at time $t_j$ on date $i$. Define

$$\Gamma = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ k \\ 1/\Delta \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 2 & \cdots & 2 & 2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2 & \cdots & k & k \\ 1/\Delta & 1/\Delta & \cdots & 1/\Delta & 1/\Delta \end{pmatrix}.$$ 

On each day $i$, the vector $Y_i = (Y_{i,t_1}, \ldots, Y_{i,t_{k+1}})'$ is normally distributed with mean 0 and covariance matrix $\sigma^2 \Delta \Sigma$. Set

$$U_{ij} = \log \left( \frac{P_{ij}}{P_{i0}} - p(t_j, Y_{ij}) \right) \quad (B.2)$$

and $U_i = (U_{i1}, \ldots, U_{i,k+1})'$. The density function of $(P_{i1}/P_{i0}, \ldots, P_{i,k+1}/P_{i0})$ conditional on $Y_i$ is

$$f(U_{i1}, \ldots U_{i,k+1}) e^{-\sum_{j=1}^{k+1} U_{ij}}.$$
where \( f \) denotes the multivariate normal density function with mean vector \(- (\delta^2 \Delta/2) \Gamma\) and covariance matrix \( \delta^2 \Delta \Sigma \).

Let \( \mathcal{L}_i \) denote the log-likelihood function for day \( i \). Dropping terms that do not depend on the parameters, we have

\[
- \mathcal{L}_i = (k + 1) \log \sigma + \frac{1}{2 \sigma^2 \Delta} Y_i' \Sigma^{-1} Y_i + (k + 1) \log \delta \\
+ \frac{1}{2 \delta^2 \Delta} \left( U_i + \frac{\delta^2 \Delta}{2} \Gamma \right)' \Sigma^{-1} \left( U_i + \frac{\delta^2 \Delta}{2} \Gamma \right) + \sum_{j=1}^{k+1} U_{ij}.
\]

Using the facts that \( \Gamma' \Sigma^{-1} = (0, \ldots, 0, 1) \) and \( \Gamma' \Sigma^{-1} \Gamma = 1 / \Delta \), this simplifies to

\[
- \mathcal{L}_i = (k + 1) \log \sigma + \frac{1}{2 \sigma^2 \Delta} Y_i' \Sigma^{-1} Y_i + (k + 1) \log \delta \\
+ \frac{1}{2 \delta^2 \Delta} U_i' \Sigma^{-1} U_i + \frac{1}{2} U_{i,k+1} + \frac{\delta^2}{8} + \sum_{j=1}^{k+1} U_{ij}.
\]

Hence, the log-likelihood function \( \mathcal{L} \) for an observation period of \( n \) days satisfies

\[
- \mathcal{L} = n(k + 1) \log \sigma + \frac{1}{2 \sigma^2 \Delta} \sum_{i=1}^{n} Y_i' \Sigma^{-1} Y_i + n(k + 1) \log \delta \\
+ \frac{1}{2 \delta^2 \Delta} \sum_{i=1}^{n} U_i' \Sigma^{-1} U_i + \frac{n \delta^2}{8} + \sum_{i=1}^{n} \left( \sum_{k=1}^{k} U_{ik} + \frac{3}{2} U_{i,k+1} \right). \quad (B.3)
\]

We minimize (B.3) in \( \alpha, \kappa, p_L, \sigma, \) and \( \delta \) (note that \( \kappa \) and \( p_L \) enter \( \mathcal{L} \) because they affect the function \( p \) that enters \( \mathcal{L} \) via (B.2)). We sample every hour and at the close, so \( \Delta = 1/6.5 \).

Appendix B.2. PIN Model

The likelihood of the PIN model is:

\[
L(B, S | \alpha, p_L, \mu, \varepsilon) = \prod_{t=1}^{T} \left\{ (1 - \alpha) \left[ \exp \left( -2 \varepsilon B_t S_t e^{B_t + S_t} \right) \right] + \alpha p_L \left[ \exp \left( - (\mu + 2 \varepsilon) B_t S_t e^{B_t + S_t} \right) \right] + \alpha (1 - p_L) \left[ \exp \left( - (\mu + 2 \varepsilon) B_t S_t e^{B_t + S_t} \right) \right] \right\} \quad (B.4)
\]
where $B_t$ ($S_t$) is the number of buys (sells) on day $t$, $\alpha$ is the probability of an information event, $p_L$ is the probability that an information event is bad news, and $\mu$ and $\varepsilon$ are the arrival rates of informed and uninformed traders. PIN, the probability of informed trade, is given by the formula:

$$PIN = \frac{\alpha \mu}{\alpha \mu + 2\varepsilon}.$$  

(B.5)
References


Li, T., 2012. Insider trading with uncertain informed trading, working Paper, City University of Hong Kong.


Table 1: Structural Parameter Correlations

Correlations of structural parameters from the hybrid and PIN models. For both models, $\alpha$ = probability of an information event and $p_L$ = probability of a negative event. For the hybrid model, $\kappa$ = signal scale parameter, $\sigma$ = standard deviation of liquidity trading, $\delta$ = volatility of public information, and Expected $\lambda$ is the expected average lambda $\lambda(0,0)$ based on Equation 8. For the PIN model, $\varepsilon$ = Poisson intensity of uninformed trades, $\mu$ = Poisson intensity of informed trades, and PIN measures the probability of informed trade.

<table>
<thead>
<tr>
<th></th>
<th>Hybrid Model</th>
<th>PIN Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$p_L$</td>
</tr>
<tr>
<td>Hybrid Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>$p_L$</td>
<td>0.04</td>
<td>1.00</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>-0.18</td>
<td>0.09</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.12</td>
<td>0.21</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.17</td>
<td>-0.09</td>
</tr>
<tr>
<td>PIN Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>-0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>$p_L$</td>
<td>-0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>-0.31</td>
<td>0.05</td>
</tr>
<tr>
<td>$\mu$</td>
<td>-0.33</td>
<td>0.10</td>
</tr>
<tr>
<td>$\nu$</td>
<td>0.23</td>
<td>-0.06</td>
</tr>
<tr>
<td>PIN</td>
<td>0.20</td>
<td>-0.02</td>
</tr>
</tbody>
</table>
Table 2: Structural Parameter Estimates and Market Capitalization

Average values of parameter estimates within each NYSE market capitalization decile (formed annually). For both models, $\alpha$ = probability of an information event and $p_L$ = probability of a negative event. For the hybrid model, $\kappa$ = signal scale parameter, $\sigma$ = standard deviation of liquidity trading, $\delta$ = volatility of public information, and Expected $\lambda$ is the expected average lambda $\lambda(0, 0)$ based on Equation 8. For the PIN model, $\varepsilon$ = Poisson intensity of uninformed trades, $\mu$ = Poisson intensity of informed trades, and PIN measures the probability of informed trade.

**Panel A. Hybrid Model**

<table>
<thead>
<tr>
<th>Decile</th>
<th>$\alpha$</th>
<th>$p_L$</th>
<th>$\kappa$</th>
<th>$\sigma$</th>
<th>$\delta$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Small)</td>
<td>0.72</td>
<td>0.57</td>
<td>0.0113</td>
<td>0.0569</td>
<td>0.0308</td>
<td>0.199</td>
</tr>
<tr>
<td>2</td>
<td>0.70</td>
<td>0.56</td>
<td>0.0081</td>
<td>0.0613</td>
<td>0.0254</td>
<td>0.144</td>
</tr>
<tr>
<td>3</td>
<td>0.67</td>
<td>0.57</td>
<td>0.0072</td>
<td>0.0706</td>
<td>0.0233</td>
<td>0.109</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
<td>0.56</td>
<td>0.0066</td>
<td>0.0798</td>
<td>0.0217</td>
<td>0.085</td>
</tr>
<tr>
<td>5</td>
<td>0.65</td>
<td>0.56</td>
<td>0.0062</td>
<td>0.0928</td>
<td>0.0205</td>
<td>0.066</td>
</tr>
<tr>
<td>6</td>
<td>0.63</td>
<td>0.57</td>
<td>0.0057</td>
<td>0.1062</td>
<td>0.0195</td>
<td>0.051</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>0.56</td>
<td>0.0055</td>
<td>0.1241</td>
<td>0.0186</td>
<td>0.043</td>
</tr>
<tr>
<td>8</td>
<td>0.59</td>
<td>0.57</td>
<td>0.0056</td>
<td>0.1512</td>
<td>0.0186</td>
<td>0.035</td>
</tr>
<tr>
<td>9</td>
<td>0.57</td>
<td>0.57</td>
<td>0.0053</td>
<td>0.1917</td>
<td>0.0177</td>
<td>0.025</td>
</tr>
<tr>
<td>10 (Large)</td>
<td>0.52</td>
<td>0.55</td>
<td>0.0062</td>
<td>0.2907</td>
<td>0.0167</td>
<td>0.020</td>
</tr>
</tbody>
</table>

**Panel B. PIN Model**

<table>
<thead>
<tr>
<th>Decile</th>
<th>$\alpha$</th>
<th>$p_L$</th>
<th>$\varepsilon$</th>
<th>$\mu$</th>
<th>$\frac{\mu}{\varepsilon}$</th>
<th>PIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Small)</td>
<td>0.31</td>
<td>0.43</td>
<td>181</td>
<td>213</td>
<td>1.53</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
<td>0.41</td>
<td>333</td>
<td>287</td>
<td>1.19</td>
<td>0.15</td>
</tr>
<tr>
<td>3</td>
<td>0.34</td>
<td>0.41</td>
<td>503</td>
<td>377</td>
<td>1.07</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>0.35</td>
<td>0.40</td>
<td>687</td>
<td>450</td>
<td>0.98</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>0.36</td>
<td>0.40</td>
<td>887</td>
<td>533</td>
<td>0.91</td>
<td>0.13</td>
</tr>
<tr>
<td>6</td>
<td>0.36</td>
<td>0.39</td>
<td>1157</td>
<td>614</td>
<td>0.85</td>
<td>0.12</td>
</tr>
<tr>
<td>7</td>
<td>0.38</td>
<td>0.38</td>
<td>1435</td>
<td>691</td>
<td>0.76</td>
<td>0.12</td>
</tr>
<tr>
<td>8</td>
<td>0.38</td>
<td>0.39</td>
<td>2000</td>
<td>837</td>
<td>0.69</td>
<td>0.11</td>
</tr>
<tr>
<td>9</td>
<td>0.39</td>
<td>0.40</td>
<td>3043</td>
<td>1031</td>
<td>0.59</td>
<td>0.09</td>
</tr>
<tr>
<td>10 (Large)</td>
<td>0.39</td>
<td>0.40</td>
<td>5210</td>
<td>1430</td>
<td>0.49</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Table 3: Price Impact and Information Asymmetry Parameters – Univariate

Univariate Fama and MacBeth (1973) cross-sectional regressions of price impact on structural parameters from the hybrid and PIN models. The structural parameters are winsorized at 1/99% and standardized to have unit standard deviation. Percent price impact is measured in basis points. Standard errors are adjusted for serial correlation following Newey and West (1987) with 5 lags. \( t \) statistics are in parentheses, and statistical significance is represented by * \( p < 0.10 \), ** \( p < 0.05 \), and *** \( p < 0.01 \).

### Panel A. Hybrid Model

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>2.72***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.20)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \kappa )</td>
<td></td>
<td>6.14***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(17.21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sigma )</td>
<td></td>
<td></td>
<td>-2.71***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-3.49)</td>
<td></td>
</tr>
<tr>
<td>Exp. ( \lambda )</td>
<td></td>
<td></td>
<td></td>
<td>4.95***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(9.27)</td>
</tr>
<tr>
<td>Constant</td>
<td>12.76***</td>
<td>12.34***</td>
<td>13.08***</td>
<td>12.84***</td>
</tr>
<tr>
<td></td>
<td>(5.08)</td>
<td>(9.80)</td>
<td>(4.76)</td>
<td>(5.67)</td>
</tr>
<tr>
<td>Observations</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.052</td>
<td>0.352</td>
<td>0.078</td>
<td>0.315</td>
</tr>
</tbody>
</table>

### Panel B. PIN Model

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>-1.67***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-3.97)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \mu )</td>
<td></td>
<td>-8.25***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-3.35)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \varepsilon )</td>
<td></td>
<td></td>
<td>-39.52**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-2.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \frac{\mu}{\varepsilon} )</td>
<td></td>
<td></td>
<td></td>
<td>5.82***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(10.89)</td>
<td></td>
</tr>
<tr>
<td>PIN</td>
<td></td>
<td></td>
<td></td>
<td>3.86***</td>
<td>(9.41)</td>
</tr>
<tr>
<td>Constant</td>
<td>13.02***</td>
<td>9.66***</td>
<td>-2.90</td>
<td>12.79***</td>
<td>12.97***</td>
</tr>
<tr>
<td></td>
<td>(4.89)</td>
<td>(12.29)</td>
<td>(-0.46)</td>
<td>(7.34)</td>
<td>(5.39)</td>
</tr>
<tr>
<td>Observations</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.037</td>
<td>0.116</td>
<td>0.143</td>
<td>0.415</td>
<td>0.200</td>
</tr>
</tbody>
</table>
Table 4: Price Impact and Information Asymmetry Parameters – Multivariate

Multivariate Fama and MacBeth (1973) cross-sectional regressions of price impact on structural parameters from the hybrid and PIN models and determinants of price impact. The structural parameters are winsorized at 1/99% and standardized to have unit standard deviation. Percent price impact is measured in basis points. In column (4), PIN is orthogonalized to Expected $\lambda$. $PIN\|\lambda$ denotes the portion collinear with Expected $\lambda$, and $PIN \perp \lambda$ denotes the orthogonal component. In columns (5) and (6), PIN is orthogonalized to the inverse of the model-implied standard deviation of order imbalances. For the PIN model, the order imbalance variance is $2\varepsilon + \alpha \mu (1 + \mu) - (\alpha \mu (1 - 2p_L))^2$. $PIN\|sd(OIB)^{-1}$ denotes the portion collinear with $sd(OIB)^{-1}$ and $PIN \perp sd(OIB)^{-1}$ denotes the orthogonal component. Each component is standardized to have unit standard deviation. Standard errors are adjusted for serial correlation following Newey and West (1987) with 5 lags. t statistics are in parentheses, and statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. $\lambda$</td>
<td>2.25***</td>
<td>1.98***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIN</td>
<td></td>
<td></td>
<td>0.66***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$PIN|\lambda$</td>
<td></td>
<td></td>
<td>2.38***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$PIN \perp \lambda$</td>
<td></td>
<td></td>
<td>0.47***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$PIN|sd(OIB)^{-1}$</td>
<td></td>
<td></td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$PIN \perp sd(OIB)^{-1}$</td>
<td></td>
<td></td>
<td>0.53**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log(Size)</td>
<td>-3.32***</td>
<td>-2.45***</td>
<td>-3.11***</td>
<td>-2.31***</td>
<td>-2.39***</td>
</tr>
<tr>
<td>Inverse Price</td>
<td>46.70***</td>
<td>50.49***</td>
<td>46.10***</td>
<td>49.91***</td>
<td>48.37***</td>
</tr>
<tr>
<td>Constant</td>
<td>58.83***</td>
<td>46.11***</td>
<td>55.84***</td>
<td>44.04***</td>
<td>46.27***</td>
</tr>
<tr>
<td>N</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
<td>20019</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.649</td>
<td>0.718</td>
<td>0.656</td>
<td>0.722</td>
<td>0.730</td>
</tr>
</tbody>
</table>
Figure 1: Price and Order Imbalance
The equilibrium price $V_t + p(t, Y_t)$ as a function of the order imbalance $Y_t$ at $t = 0.5$ for two values of the probability $\alpha$ of an information event, when $\xi_S = H$, $V_t = 50$, $H = 10$, $L = -10$, $\sigma = 1$, and $p_H = p_L = 1/2$. 

![Graph showing price and order imbalance relationship](image-url)
Figure 2: Expected Average Lambda
Expected average lambda (8) as a function of $\alpha$ for two different values of $H - L$, taking $|L| = H$, when $\sigma = 1$ and $p_L = p_H = 1/2$. 
Figure 3: Time Series of Hybrid Model Estimates
The annual cross-sectional mean and 25th and 75th percentiles of parameter estimates for the hybrid model. The model is estimated on a stock-year basis for NYSE stocks from 1993 through 2012 using prices and order imbalances in six hourly intraday bins and at the close. The model parameters are $\alpha =$ probability of an information event, $\kappa =$ signal scale parameter, $\sigma =$ standard deviation of liquidity trading, $\delta =$ volatility of public information, and $p_L =$ probability of a negative event. Expected $\lambda$ is the expected average lambda $\lambda(0,0)$ based on Equation 8.

(a) $\alpha$  
(b) $\kappa$  
(c) $p_L$  
(d) $\sigma$  
(e) $\delta$  
(f) Expected $\lambda$
Figure 4: Time Series of PIN Model Estimates
The annual cross-sectional mean and 25th and 75th percentiles of parameter estimates for the Easley et al. (1996) model. The model is estimated on a stock-year basis for NYSE stocks from 1993 through 2012 using daily buys and sells. The model parameters are $\alpha =$ probability of an information event, $p_L =$ probability of a negative event, $\varepsilon =$ Poisson intensity of uninformed trades, $\mu =$ Poisson intensity of informed trades, and PIN = Probability of informed trade.

(a) $\alpha$

(b) $p_L$

(c) $\varepsilon$

(d) $\mu$

(e) $\mu / \varepsilon$

(f) PIN
Figure 5: Time Series of Price Impact Estimates
The annual cross-sectional mean and 25th and 75th percentiles of percent price impacts. Five-minute price impacts are estimated daily and averaged annually for each stock-year for NYSE stocks from 1993 through 2012.
Figure 6: The equilibrium informed trading rate $\theta_t$ as a function of the price $V_t + p(t,Y_t)$ at $t = 0.5$ for two values of the probability $\alpha$ of an information event, when $\xi S = H$, $V_t = 50$, $H = 10$, $L = -10$, $\sigma = 1$, and $p_H = p_L = 1/2$. 
Figure 7: Conditional Order Imbalance Distributions
The density function of the net order flow $Y_1$ conditional on a low signal, no information event, and a high signal for two values of the probability $\alpha$ of an information event, when $\sigma = 1$ and $p_L = p_H = 1/2$. 
Figure 8: Model-implied Order Imbalance Distributions and Market Capitalization
The mixture distributions of standardized order imbalances implied by structural estimates from the hybrid and PIN models for the smallest and largest size deciles. Order imbalances are standardized by the standard deviation of order imbalances. For the hybrid model, the order imbalance variance is \( \sigma^2 \). For the PIN model, the order imbalance variance is \( 2\varepsilon + \alpha\mu(1 + \mu) - (\alpha\mu(1 - 2p_L))^2 \). The parameters for each size decile are based on the structural estimates reported in Table 2.

(a) Smallest Size Decile (Hybrid)  (b) Largest Size Decile (Hybrid)

(c) Smallest Size Decile (PIN)  (d) Largest Size Decile (PIN)
Figure 9: Empirical Order Imbalance Distributions and Market Capitalization
The distributions of daily standardized order imbalances for the smallest and largest size deciles. For each firm-year, daily order imbalances are standardized by the firm-year standard deviation. The hybrid model is estimated using order imbalances measured in shares (top row) and the PIN models are estimated using order imbalances measured in number of trades (bottom row).

(a) Smallest Size Decile (Empirical Shares)       (b) Largest Size Decile (Empirical Shares)

(c) Smallest Size Decile (Empirical Trades)      (d) Largest Size Decile (Empirical Trades)