# Is Information Risk a Determinant of Asset Returns?

By

David Easley
Department of Economics
Cornell University

Soeren Hvidkjaer
Johnson Graduate School of Management
Cornell University

Maureen O'Hara
Johnson Graduate School of Management
Cornell University

June 2000

* The authors would like to thank Kerry Back, Patrick Bolton, Douglas Diamond, Ken French, William Gebhardt, Mark Grinblatt, Campbell Harvey, David Hirshleifer, Schmuel Kandell, Charles Lee, Bhaskaran Swaminathan, Zhenyu Wang, Ingrid Werner, two referees and the editor (Rene Stulz), and seminar participants at Cornell University, DePaul University, the Federal Reserve Bank of Chicago, Georgia State University, the Hong Kong University of Science and Technology, INSEAD, the Massachusetts Institute of Technology, Princeton University, the Red Sea Finance Conference, the Stockholm School of Economics, Washington University, the University of Chicago, and the Western Finance Association meetings for helpful comments. We are also grateful to Marc Lipson for providing us with data compaction programs. Please direct comments and correspondence to mo19@cornell.edu

# Is Information Risk a Determinant of Asset Returns?

## Abstract

In this research we investigate the role of information-based trading in affecting asset returns. Our premise is that in a dynamic market asset prices are continually adjusting to new information. This evolution dictates that the process by which asset prices become informationally efficient cannot be separated from the process generating asset returns. Using the structure of a sequential trade market microstructure model, we derive an explicit measure of the probability of information-based trading for an individual stock, and we estimate this measure using high-frequency data for NYSE-listed stocks for the period 1983-1998. The resulting estimates are a time-series of individual stock probabilities of information-based trading for a very large cross section of stocks. We investigate whether these information probabilities affect asset returns by incorporating our estimates into a Fama-French [1992] asset pricing framework. Our main result is that information does affect asset prices: stocks with higher probabilities of information-based trading require higher rates of return. Indeed, we find that a difference of 10 percentage points in the probability of information-based trading between two stocks leads to a difference in their expected returns of 2.5% per year. We interpret our results as providing strong support for the premise that information affects asset pricing fundamentals.

# Is Information Risk a Determinant of Asset Returns?

## 1. Introduction

Asset pricing is fundamental to our understanding of the wealth dynamics of an economy. This central importance has resulted in an extensive literature on asset pricing, much of it focusing on the economic factors that influence asset prices. Despite the fact that virtually all assets trade in markets, one set of factors not typically considered in asset pricing models are the features of the markets in which the assets trade. Instead, the literature on asset pricing abstracts from the mechanics of asset price evolution, leaving unsettled the underlying question of how equilibrium prices are actually attained.

Market microstructure, conversely, focuses on how the mechanics of the trading process affect the evolution of trading prices. A major focus of this extensive literature is on the process by which information is incorporated into prices. The microstructure literature provides structural models of how prices become efficient, as well as models of volatility, both issues clearly of importance for asset pricing. But of perhaps more importance, microstructure models can provide explicit estimates of the extent of private information. The microstructure literature has demonstrated the important link between this private information and an asset's bid and ask trading prices, but it has yet to be demonstrated that such information actually affects asset pricing fundamentals.

If a stock has a higher probability of private information-based trading, should that have an effect on its required return? In traditional asset pricing models, the answer is no. These models rely on the notion that if assets are priced "efficiently", then information is already incorporated and hence need not be considered. But this view of efficiency is static, not dynamic. If asset prices are continually revised to reflect new information, then efficiency is a process, and how asset prices become efficient cannot be separated from asset returns at any point in time.

This issue of information and asset returns has been addressed in various ways in the literature. Perhaps the most straightforward approach is that of Amihud and Mendelson [1986] who consider a variant of this problem by arguing that liquidity should be priced. Their argument is that investors maximize expected returns net of trading costs, proxied by the bid-ask spread. Therefore, in equilibrium, higher returns are required for stocks with higher spreads. Amihud and Mendelson [1986; 1989] and Eleswarapu [1997] present empirical evidence consistent with this liquidity hypothesis. Supporting evidence using other measures of liquidity is provided by Amihud, Mendelson and Lauterbach [1998], Amihud [2000], Datar, Naik, and Radcliffe [1998], Brennan and Subrahmanyam [1996], and Brennan, Chordia and Subrahmanyam [1998]. But the overall research on this issue is mixed, with Chen and Kan [1996], Eleswarapu and Reinganum [1993], and Chalmers and Kadlec [1998] concluding that liquidity is not priced. Certainly, one might agree with Datar, Naik, and Radcliffe's observation that "whether liquidity affects asset returns or not remains unresolved thus far".

One difficulty in resolving this issue lies in what exactly is being sought. Is this higher return, if it exists, due to a compensation for some exogenous illiquidity that manifests itself in large spreads? Or is it a return for bearing the risk of trading with counterparties who have superior information, a factor that would also induce high spreads? Illiquidity and information risk are obviously related issues, but they are not the same. The illiquidity arising from some exogenous factors (such as limited competition between dealers), is akin to a tax, and its effects might be reasonably anticipated as a positive link between spreads and returns. The effects of private information are more complex, however, because of their link to the dynamic efficiency of asset prices. Do traders need compensation to hold a stock that has a greater risk of information-based trading?

The distinction between these two concepts can be illustrated by a simple example. Consider an investor choosing between investments in two stocks. Suppose that the two stocks are identical in every way except that in one stock all information events are public and in the other all information events are private. The stock with private information events will, according to standard market microstructure models, have a larger spread than the stock with public information events. But this is surely only a minor concern for the investor. Of more importance is that the stock with private information events is riskier for the uninformed investor than is the stock in which the events are public. The uninformed investor must be rewarded in equilibrium for the risk of holding this stock, and we argue here that it is this information risk that is priced in asset returns.

Our focus in this paper is on showing empirically that information risk affects cross-sectional asset returns. We first present a simple model to provide the intuition for why private information affects stock returns.[1] We then develop an empirical methodology for estimating this effect by incorporating an explicit microstructure measure of information-based trading into an asset-pricing framework. Our analysis uses a structural market microstructure model to generate a measure of the probability of information-based trading (PIN) in an individual stock. We then estimate this measure using high-frequency data for NYSE-listed stocks for the period 1983-1998. The resulting estimates are a time-series of individual stock probabilities of information-based trading for a very large cross section of stocks. We investigate whether these information probabilities affect cross-sectional asset returns by incorporating our estimates into a Fama-French [1992] asset-pricing framework. Our main result is that information does affect asset prices: stocks with higher probabilities of information-based trading have higher rates of return. Indeed, we find that a difference of 10 percentage points in PIN between two stocks leads to a difference in their expected returns of 2.5% per year. The magnitude, and statistical significance, of this effect provides strong support for the premise that information affects asset-pricing fundamentals.

Our focus on the role of information in asset pricing is related to several recent papers. In a companion theoretical paper, Easley and O'Hara [2000] develop a multi-asset rational expectations equilibrium model in which stocks have differing levels of public and private information. In equilibrium, uninformed traders require compensation to hold stocks with greater private information, resulting in cross-sectional differences in returns. The basic intuition of this model is outlined in the next section, and it forms the basis for our empirical estimation. Wang [1993] provides an intertemporal asset-pricing model in which traders can invest in a risk-less asset and a single risky asset. In this model, the presence of traders with superior information induces an adverse selection problem, as uninformed traders demand a premium for the risk of trading with informed traders. However, trading by the informed investors also makes prices more informative, thereby reducing uncertainty. These two effects go in opposite directions, and their overall effect on asset returns is ambiguous. Because this model allows only one risky asset, it is not clear how, if at all, information would affect cross-sectional returns. Jones and Slezak [1999] also develop a theoretical model allowing for asymmetric information to affect asset returns. Their model relies on changes in the variance of news and liquidity

[1]The theoretical case for why information affects asset returns is developed more fully in Easley and O'Hara [2000]. We present in this paper a brief theoretical explanation of why this cross-sectional effect arises.

shocks over time to differentially affect agents' portfolio holdings, thereby influencing asset returns. These theoretical papers suggest that information can affect asset returns, the issue of interest in this paper.

Two recent empirical papers related to our analysis are Brennan and Subrahmanyam [1996] and Amihud [2000]. These authors investigate how the slope of the relation between trade volume and price changes affects asset returns. This measure of illiquidity relies on the price impact of trade, and it seems reasonable to believe that stocks with a large illiquidity measure are less attractive to investors. Brennan and Subrahmanyam find support for this notion using 2 years of transactions data to estimate the slope coefficient $\lambda$, while Amihud establishes a similar finding using daily data. What economic factors underlie this result is not clear. Because $\lambda$ is derived from price changes, factors such as the impact of price volatility on daily returns, or inventory concerns by the market maker could influence this variable, as could adverse selection.[2] Neither analyses addresses whether their illiquidity measure is proxying for spreads, or for the more fundamental information risk we address. Our analysis here focuses directly on private information by deriving a trade-based measure of information risk. This PIN measure has been shown in previous work (see Easley, Kiefer, and O'Hara [1996; 1997a; 1997b], Easley, Kiefer, O'Hara and Paperman [1996], and Easley, O'Hara and Paperman [1998]) to explain a number of information-based regularities, providing the link to private information we need to investigate cross-sectional asset pricing returns.

The PIN variable is correlated with other variables that we do not include in our return estimation. In particular, as would be expected with an information measure, PIN is correlated with spreads. It is also correlated with the variability of returns and with volume or turnover. One might suspect that the probability of information based trade only seems to be priced because it serves as a proxy for these omitted variables. We show, however, that this is not the case. We show that over our sample period, spreads do not affect asset returns but PIN does. When spreads or the variability of returns are included with PIN in the return regressions, the

[2] The Kyle $\lambda$ has not been tested as to its actual linkage with private information. While it seems reasonable to us that such a theoretical linkage would exist, there are a number of reasons why this empirical measure is problematic. For example, the actual Kyle model assumes a call market structure in which orders are aggregated and it is only the net imbalance that affects the price. Actual markets do not have this structure, so in practice $\lambda$ is estimated on a trade-by-trade basis (as in BS), or is a time series change in price per volume over some interval (as in Amihud). Either approach may introduce noise in the specification. Moreover, because the $\lambda$ calculation also involves both price and the quantity of the trade its actual value may be affected by factors such as the size of the book , tick size consideration, and market maker inventory.

probability of information based trade remains highly significant, and its effect on returns is changed only slightly. Volume remains a factor in asset pricing, but it does not remove the influence of PIN. We view these results as strong evidence that the probability of information based trade is priced in asset returns.

The paper is organized as follows. Section 2 provides the theoretical intuition for our analysis by outlining a rational expectation model in which traders receive both public and private information signals about a number of stocks. This model demonstrates that private information affects asset returns because it skews the portfolio holdings of informed and uninformed traders in equilibrium. We then turn in Section 3 to the empirical testing methodology. We set out a basic microstructure model and we demonstrate how the probability of information-based trading is derived for a particular stock. Estimation of the model involves maximum likelihood, and we show how to derive these estimating equations. In Section 4 we present our estimates. We examine the cross-sectional distribution of our estimated parameters, and we examine their temporal stability. A simple check on the reasonableness of our estimates of information-based trading is to examine their relation to opening spreads. We find that our model does a very good job of explaining spreads, and we find the independently interesting result that spreads experienced a structural shift following the 1987 crash. Section 5 then puts our estimates into an asset pricing framework. We use the cross-sectional approach of Fama-French [1992] to investigate expected asset returns. In this section we present our results, and we investigate their robustness. We also investigate the differential ability of spreads, variability of returns, turnover, and our information measure to affect returns. The paper's last section summarizes our results and discusses their implications for asset pricing research.

## 2. **Information and Asset Prices**

To show why trading based on private information should affect asset returns, we construct a simple rational expectations equilibrium asset-pricing model. We use this analysis only to motivate our empirical search for information effects so we keep the exposition here as simple as possible. A complete model deriving the rational expectations equilibrium and investigating the specific effects of public and private information on asset prices is found in a companion theoretical paper Easley and O'Hara [2000].

We consider a two-period model: today when investors choose portfolios and tomorrow when the assets in these portfolios payoff. There is one risk free asset, money, which has a constant price of 1. There are K risky stocks indexed by k=1,…,K. The future value, $v_k$, of stock k is random with distribution $N(\overline{v_k}, \rho_k^{-1})$. We let $p_k$ denote the price today of a share of stock k. There are signals that some or all investors will receive today about the future values of these stocks. For stock k, $I_k$ signals are drawn independently from the distribution $N(v_k, \gamma_k^{-1})$. Some of these signals are public and some are private. The fraction of the $I_k$ signals about the value of stock k that are private is denoted $\alpha_k$; the fraction of signals that are public is 1- $\alpha_k$. All investors receive any public signals before trade begins. Only informed traders receive any private signals. We let $\mu_k$ be the fraction of traders who receive the private signals about stock k. Finally the aggregate supply of shares of stock k is random with distribution $N(\overline{x_k}, \eta_k^{-1})$ with $\overline{x}_k > 0$. All random variables are independent, and their distributions are known to the investors.

There are K+1 assets, hence K relative prices, and many sources of uncertainty: signals about the future value of the stocks and the random supply of each stock. We view the random supply of stocks as a simple proxy for noise trade, but it is important. Without the high dimensional information space there would be a fully revealing rational expectations equilibrium in which the uninformed investors could completely infer the informed investors' information from equilibrium asset prices. It would then not matter whether information was public or private.

There are J investors indexed by j=1,…, J. These investors all have CARA utility with coefficient of risk aversion $\delta$. Investors are endowed with money; $\overline{m^j} > 0$. These investors must in equilibrium hold the available supply of money and stocks. Markets are incomplete, so stocks are risky even for informed investors. Because the investors are risk averse, and the stocks are risky, the risk will be priced in equilibrium. The question that we are interested in is how the distribution of information affects asset prices and thus expected returns.

The budget constraint today for typical investor j is $m^j + \sum_k p_k z_k^j = \overline{m^j}$, where $z_k^j$ is the number of shares of stock k he purchases and $m^j$ is the amount of money he holds. His wealth tomorrow is the random variable $w^j = \sum_k v_k z_k^j + m^j$. Suppose that conditional on all of investor j's information his predicted distribution of $v_k$ is $N(\overline{v}_k^j, (\rho_k^j)^{-1})$. Then his optimal demand for stock k is given by

(1) $$z_k^j = \frac{\overline{v_k^j} - p_k}{\delta(\rho_k^j)^{-1}}.$$

Thus the equilibrium price of stock *k* is

(2) $$p_k = \frac{\sum_j \overline{v_k^j}\rho_k^j - \delta x_k}{\sum_j \rho_k^j}.$$

Computation of equilibrium prices requires showing that for both informed and uninformed investors conditional distributions are Normal. This is trivial for informed investors. It is less trivial for uninformed investors because of the inferences that they draw from equilibrium prices, but it is nonetheless true in at least one linear equilibrium (see Easley and O'Hara [2000] for derivation). In this rational expectations equilibrium the (prior) expected excess return on stock *k* is

(3) $$E[v_k - p_k] = \frac{\delta \overline{x_k}}{\rho_k + (1-\alpha_k)I_k\gamma_k + (1-\mu_k)\alpha_k I_k \theta_k},$$

where $\theta_k = \left[(\mu_k\gamma_k)^{-2}(\alpha_k I_k)^{-1}\eta_k^{-1}\delta^2 + \gamma_k^{-1}\right]^{-1}$ is the precision of the uninformed traders' posterior distribution on the value of stock k.

Equation (3) provides the rationale for why private information affects equilibrium asset prices. If agents are risk averse (δ>0), and if stock *k* is in positive net supply on average ($\bar{x}_k > 0$), then its price must on average be less than its expected future value. This is because in equilibrium risk averse investors must be compensated for holding the positive supply of the stock. Information affects this return because it affects the risk of holding the asset. If there is perfect prior information ($\rho_k = \infty$) or perfect signals ($\gamma_k = \infty$) then all traders know the asset's true value, so it is risk free and its price is its expected future value. In a risk free or fully revealing equilibrium all traders hold the same portfolio of assets. Otherwise, in equilibrium the informed hold more of the good news stocks and less of the bad news stocks, necessitating a risk premium to induce the uninformed to hold the risky assets. Calculation shows that if private signals are truly private ($\mu_k$<1), then the expected excess return is increasing in $\alpha_k$, the fraction of the signals about stock k that are private.

This result provides our main hypothesis: in comparing two stocks that are otherwise identical, the stock with more private and less public information will have a larger expected excess return. This occurs because when information is private, rather than public, uninformed investors cannot perfectly infer it from prices, and they consequently view the stock as being more risky.

Uninformed investors could avoid this risk, but they chose not to do so. To completely avoid this risk the uninformed traders would have to hold only money, but this is not optimal; they receive higher utility by holding some of the risky stocks. They are rational, so they hold an optimally diversified portfolio, but no matter how they diversify they are taken advantage of by the informed traders who know better which stocks to hold. Although the model has only one trading period, it is easy to see that uninformed investors also would not chose to avoid this risk by buying and holding a fixed portfolio over time. In each trading period in an inter-temporal model uninformed investors reevaluate their portfolios. As prices change, they optimally change their holdings.

The model demonstrates that the extent of private versus public information affects equilibrium asset returns, but testing it requires a mechanism for measuring information-based trading.[3] This measure can be derived from a market microstructure model, and it is to this derivation that we now turn.

## 3. Microstructure and Asset Prices

Consider what we know from the microstructure literature (see O'Hara [1995] for a discussion and derivation of microstructure models). Microstructure models can be viewed as learning models in which market makers watch market data and draw inferences about the underlying true value of an asset. Crucial to this inference problem is their estimate of the probability of trade based on private information about the stock. Market makers watch trades,

[3] If a stock has more private information and an unchanged amount of public information its equilibrium expected return falls. This occurs because risk is reduced. Here we keep the underlying information structure fixed and vary the split of this information between public and private.

update their beliefs about this private information, and set trading prices. Over time, the process of trading, and learning from trading, results in prices converging to full information levels.

As an example, consider the simple sequential trade tree diagram given in Figure 1. Microstructure models depict trading as a game between the market maker and traders that is repeated over trading days i=1,…,I. First, nature chooses whether there is new information at the beginning of the trading day, these events occur with probability $\alpha$. The new information is a signal regarding the underlying asset value, where good news is that the asset is worth $\overline{V}_i$, and bad news is that it is worth $\underline{V}_i$. Good news occurs with probability (1-$\delta$) and bad news occurs with the remaining probability, $\delta$. Trading for day i then begins with traders arriving according to Poisson processes throughout the day. The market maker sets prices to buy or sell at each time t in [0,T] during the day, and then executes orders as they arrive. Orders from informed traders arrive at rate $\mu$ (on information event days), orders from uninformed buyers arrive at rate $\varepsilon_b$ and orders from uninformed sellers arrive at rate $\varepsilon_s$. Informed traders buy if they have seen good news and sell if they have seen bad news. If an order arrives at time t, the market maker observes the trade (either a buy or a sale), and he uses this information to update his beliefs. New prices are set, trades evolve, and the price process moves in response to the market maker's changing beliefs. This process is captured in Figure 1.

Now suppose we view this problem from the perspective of an econometrician. If we, like the market maker, observed a particular sequence of trades, what could we discover about the underlying structural parameters and how would we expect prices to evolve? This is the intuition behind a series of papers by Easley, Kiefer, and O'Hara (1996; 1997a; 1997b) who demonstrate how to use a structural model to work backwards to provide specific estimates of the risks of information-based trading in a stock. They show that these structural models can be estimated via maximum likelihood, providing a method for determining the probability of information-based trading in a given stock. In particular, the likelihood function induced by this simple model of the trade process for a single trading day is

(4) $$\begin{aligned} L(\theta \mid B,S) &= (1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \\ &+ \alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(\mu+\varepsilon_s)}\frac{(\mu+\varepsilon_s)^S}{S!} \\ &+ \alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \end{aligned}$$

where B and S represent total buy trades and sell trades for the day respectively, and $\theta = (\alpha, \mu, \varepsilon_B, \varepsilon_S, \gamma)$ is the parameter vector. This likelihood is a mixture of distributions where the trade outcomes are weighted by the probability of it being a "good news day" $\alpha(1-\delta)$, a "bad news day" $(\alpha\delta)$, and a "no-news day" $(1-\alpha)$.

Imposing sufficient independence conditions across trading days gives the likelihood function across I days

(5) $$V = L(\theta \mid M) = \prod_{i=1}^{I} L(\theta \mid B_i, S_i)$$

where $(B_i, S_i)$ is trade data for day i = 1,…,I and $M=((B_1,S_1), \ldots,(B_I,S_I))$ is the data set.[4] Maximizing (5) over $\theta$ given the data M thus provides a way to determine estimates for the underlying structural parameters of the model ( i.e. $\alpha$, $\mu$, $\varepsilon_B$, $\varepsilon_S$, $\delta$).

This model allows us to use observable data on the number of buys and sells per day to make inferences about unobservable information events and the division of trade between the informed and uninformed. In effect, the model interprets the normal level of buys and sells in a stock as uninformed trade, and it uses this data to identify $\varepsilon_B$ and $\varepsilon_S$. Abnormal buy or sell volume is interpreted as information-based trade, and it is used to identify $\mu$. The number of days in which there is abnormal buy or sell volume is used to identify $\alpha$ and $\delta$. Of course, the maximum likelihood actually does all of this simultaneously. For example, consider a stock that always has 40 buys and 40 sells per day. For this stock, $\varepsilon_B$ and $\varepsilon_S$ would be identified as 40 (where the parameters are daily arrival rates), $\alpha$ would be identified as 0, and $\delta$ and $\mu$ would be

[4] The independence assumptions essentially require that information events are independent across days. Easley, Kiefer, and O'Hara [1997b] do extensive testing of this assumption and are unable to reject the independence of days.

unidentified. Suppose, instead, that on 20% of the days there are 90 buys and 40 sells; and, on 20% of the days there are 40 buys and 90 sells. The remaining 60% of the days continue to have 40 buys and 40 sells. The parameters in this example would be identified as $\varepsilon_B = \varepsilon_S=40$, $\mu=50$, $\alpha=0.4$ and $\delta=0.5$.

One might conjecture that this trade imbalance statistic is too simplistic to capture the actual influence of informed trading. In particular, because trading volume naturally fluctuates, perhaps these trade imbalance deviations are merely natural artifacts of random market influences, and are not linked to information-based trade as argued here. However, it is possible to test for this alternative by restricting the weights on the mixture of distributions to be the same across all days. This "random volume" model is soundly rejected in favor of information-mixture derive above (see Easley, Kiefer, and O'Hara [1997b] for procedure and estimation results). A second concern is that the model uses only patterns in the number of trades, and not patterns in volume, to identify the structural parameters.[5] It is possible to add trade size to the underlying approach, in which case the sufficient statistic for the trade process is the four-tuple (#large buys, #large sells, #small buys, and #small sales). This greatly increases the computational complexity, but as shown in Easley, Kiefer, and O'Hara [1997a], there appears to be little gain in doing so as the trade size variables do not generally reveal differential information content. Given the extensive estimation required in this project, we have chosen to use the simple model derived above; to the extent that this omits important factors, we would expect the ability of our estimates to predict asset pricing returns to be reduced.

We now turn to the economic use of our structural parameters. The estimates of the model's structural parameters can be used to construct the theoretical opening bid and ask prices.[6] As is standard in microstructure models, a market maker sets trading prices such that his expected losses to informed traders just offset his expected gains from trading with uninformed traders. This balancing of gains and losses is what gives rise to the "spread" between bid and ask prices. As demonstrated in the Appendix, for the five parameter model analyzed here, the model predicts the percentage opening spread on day i to be

---

[5] This number of trades approach is consistent with the findings of Jones, Kaul and Lipson [1994], who argue that volume does not provide information beyond number of trades.

(6) $$\frac{\Sigma}{V*_i} = \mu\sqrt{\alpha\delta(1-\delta)}\left[\frac{1}{\varepsilon_b + \mu\alpha(1-\delta)} + \frac{1}{\varepsilon_s + \mu\alpha\delta}\right]\sigma_v$$

where $V*_i$ is the unconditional expected value of the asset given by $V*_i = \delta\underline{V}_i + (1-\delta)\overline{V}_i$, and $\sigma_v$ is the standard deviation of the daily percentage price change. Intuitively, this equation yields some natural, and economically reasonable, comparative statics: the higher the fraction of informed traders ($\mu$) or the more likely are information events ($\alpha$), the greater is the spread; the greater the arrival rates of uninformed orders ($\varepsilon_B$ and $\varepsilon_S$) the smaller is the spread. The standard deviation enters because the market maker's expected losses are higher the greater the divergence in potential asset prices. An important feature to note is that the absence of new information ($\alpha$) or traders informed of it ($\mu$), results in a zero spread. This reflects the fact that only asymmetric information affects spreads when market makers are risk neutral.

The opening spread is easiest to interpret if we express it explicitly in terms of this information-based trading. It is straightforward to show that the probability that the opening trade is information-based, PIN, is

(7) $$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_S + \varepsilon_B}$$

where $\alpha\mu + \varepsilon_S + \varepsilon_B$ is the arrival rate for all orders and $\alpha\mu$ is the arrival rate for information-based orders. The ratio is thus the fraction of orders that arise from informed traders or the probability that the opening trade is information-based. In the economically sensible case in which the uninformed are equally likely to buy and sell ($\varepsilon_b = \varepsilon_s = \varepsilon$) and news is equally likely to be good or bad ($\delta = 0.5$), the percentage opening spread equation in equation (6) simplifies to

(8) $$\frac{\Sigma}{V*_i} = (PIN)\frac{(\overline{V}_i - \underline{V}_i)}{V*_i}$$

---

[6] Given any history of trades we can also construct the theoretical bid and ask prices at any time during the trading day. But in our empirical work we focus on opening prices so we provide here only the derivation for the opening spread.

Returning to our example of a stock that has trade resulting in estimated parameters of $\varepsilon_B = \varepsilon_S$=40, $\mu$=50, $\alpha$=0.4 and $\delta$=0.5, we see that PIN for this stock would be 0.2. This means that for this stock the market maker believes that 20% of the trades come from informed traders. This risk of information based trade results in a spread , but the size of this spread also depends on the variability of the value of the stock. If this stock typically has a range of true values of $4 around an expected value on day i of $50 then its opening spread, $\Sigma$, would be predicted to be $0.80 resulting in an opening percentage spread around $50 of 1.6%.

Neither the estimated measure of information-based trading nor the predicted spread is related to market maker inventory because these factors do not enter into the model. Instead, these estimates represent a pure measure of the risk of private information. More complex models can also be estimated, allowing for greater complexity in the trading and information processes. Easley, Kiefer, and O'Hara [1996; 1997a; 1997b], Easley, Kiefer, O'Hara and Paperman [1996], and Easley, O'Hara and Paperman [1998] have used these measures of asymmetric information to show how spreads differ between frequently and infrequently traded stocks, to investigate how informed trading differs between market venues, to analyze the information content of trade size, and to determine if financial analysts are informed traders.

Whether asymmetric information also affects required asset returns is the issue of interest in this paper. The model and estimating procedure detailed above provide a mechanism for determining the probability of information-based trading, and it is this PIN variable that we will explore in an asset pricing context in Section 5 of this paper. Asset pricing considerations, however, are inherently dynamic, focussing as they do on the return that traders require over time to hold a particular asset. This dictates that any information-linked return must also be dynamic, and hence we need to focus on the time-series properties of our estimated information measure. Prefatory to this, however, is the more fundamental problem of estimating PIN when the underlying structural variables can be time-varying.

In the next section we address these estimation issues. Using time series data for a cross section of stocks, we maximize the likelihood functions given by our structural model. We use our estimates of the structural parameters to calculate PIN, and we investigate the temporal stability of these estimates. A fundamental difficulty in any empirical investigation is determining whether the estimates actually measure what they purport to measure. That is, since the probability of information-based trading is inherently unobservable, a natural concern is that our estimates do not actually capture the underlying asymmetric information. We address this concern by examining how well our estimates do in explaining spread behavior. It is generally agreed that information-based trading affects spreads, and so we test the economic properties of our estimated spread (6) using both cross-sectional actual spreads and the time-series of actual spreads. Having established the statistical properties and economic validity of our estimates, we then address the link between information and asset-pricing in the following section.

## 4. **The Estimation of Information-based Trading**

### *4.1 Data and Methodology*

We estimate our model for a sample of all ordinary common stocks listed on the New York Stock Exchange for the years 1983-1998. We focus on NYSE-listed stocks because the market microstructure of that venue most closely conforms to that of our structural model. We exclude REITS, stocks of companies incorporated outside of the U.S, and closed end funds. We also exclude a stock in any year in which it did not have at least 60 days with quotes or trades, as we cannot estimate our trade model reliably for such stocks. This leaves us with a sample of between 1311 and 1846 stocks to be analyzed each year.

The likelihood function given in equation (5) depends upon the number of buys and sells each day for each stock in our sample. Transactions data gives us the daily trades for each of our stocks, but we need to classify these trades as buys or sells. To construct this data, we first retrieve transactions data from the Institute for the Study of Security Markets (ISSM) and Trade And Quote (TAQ) datasets. We then classify trades as buys or sells according to the Lee-Ready algorithm (see Lee and Ready [1991]). This algorithm is standard in the literature and it essentially uses trade placement relative to the current bid and ask quotes to determine trade direction.[7] Using this data, we maximize the likelihood function over the structural parameters, $\theta = (\alpha, \mu, \varepsilon_{B,} \varepsilon_S, \delta)$, for each stock separately for each year in the sample period. This gives us one yearly estimate per stock for each of the underlying parameters.[8]

---

[7] See Ellis, Michaely, and O'Hara [1999] for an analysis of alternative trade classification algorithms and their accuracy.

[8] We chose an annual estimation period because of the need to estimate the time series of the large number of stocks in our sample. The model can be estimated using as little as 60 trading days of data provided there is sufficient trading activity. We estimated our parameters over rolling 60-day windows for a sub-sample of stocks, but found little difference with the annual estimates.

The underlying model involves two parameters relating to the daily information structure ($\alpha$, the probability of new information, and $\delta$, the probability that new information is bad news) and three parameters relating to trader composition ($\mu$, the arrival rate of informed traders, and $\varepsilon_S$ and $\varepsilon_b$, the arrival rates of uninformed buyers and sellers). Information on $\mu$, $\varepsilon_S$ and $\varepsilon_b$ accumulates at a rate approximately equal to the square root of the number of trade outcomes, while information on $\alpha$ and $\delta$ accumulates at a rate approximately equal to the square root of the number of trading days. The difference in information accumulation rates dictates that the precision of our $\mu$ and $\varepsilon$ estimates will exceed that of our $\alpha$ and $\delta$ estimates, but the length of our time series is more than sufficient to provide precise estimates of each variable.

The maximum likelihood estimation converges for almost all stocks. Of more than 20,000 time series, only 716 did not converge. These failures were generally due to series with days of such extremely high trading volume compared to normal levels that convergence was not possible. Further, the estimation yielded only 456 corner solutions in $\delta$, the probability of an information event being bad news. Such corner solutions arise because a sustained imbalance of trading (e.g. more buys than sells) will result in the estimates of the probability of bad news being driven to one or zero. There are only 6 corner solutions found for $\alpha$, the probability of any day being an information day.[9] This finding is reassuring as it suggests the economically reasonable result that private information is a factor in the trading of every stock.

*4.2 Distribution of Parameter Estimates*

The time series patterns of the cross sectional distribution of the individual parameter estimates are shown in Figure 2. The parameter estimates generally exhibit reasonable economic behavior. The estimates of $\mu$ , $\varepsilon_S$ and $\varepsilon_b$ are related to trading frequency, and hence show an upward trend as trading volume increases on the NYSE over our sample period.[10] On the other hand, the estimates of $\alpha$ and $\delta$ are stable across years, and so, as expected, they do not trend.

Our particular interest is in the composite variable PIN, the probability of information-based trading. PIN is computed from equation (7) using the yearly estimates of $\alpha$, $\delta$, $\mu$ , $\varepsilon_S$ and

---

[9] The better performance of $\alpha$ over $\delta$ is not surprising, as only the fraction of days that have information events is used for the estimation of $\delta$, while the algorithm uses the whole sample in estimating $\alpha$. Indeed, corner solutions to $\delta$ are mainly found in stocks with low $\alpha$ estimates.

[10] These estimates also show a peak at the time of the 1987 market crash, and a fall-off in the low volume years following the crash.

$\varepsilon_b$, thus we obtain one estimate of PIN for each stock each year. The estimated PIN is very stable across years, both individually and cross-sectionally. Panel A of Figure 3 shows the cross-sectional pattern of PIN. Not only is the median almost constant around 0.19, but the individual percentiles also appear to be stable across years. On an individual stock level, absolute changes between years are relatively small. Panel B of Figure 3 shows the cumulative distribution of year-to-year absolute changes in individual stock PIN. We find that 50% of absolute changes are within 3 percentage points (out of a median of 19 percentage points), while 95% are within 11 percentage points. Thus, individual stocks exhibit relatively low variability in the probability of information-based trading across years.

An interesting question is how these PIN estimates relate to the underlying trading volume in the stock. We calculated the cross-sectional correlations between PIN and the logarithm of average daily trading volume for each stock for each year of our sample. The average correlation over the 16 years in our sample is –0.54, with a range of –0.38 to –0.71. Hence, we find that across stocks within the same year, PIN is negatively correlated with trading volume. This is consistent with previous empirical work (see Easley, Kiefer, O'Hara and Paperman [1996]) showing that actively traded stocks face a lower adverse selection problem due to informed trading. Note then that across stocks within a each year PIN is negatively correlated with trading volume, while across time, PIN estimates remain constant, even though trading volume increases. These are exactly the patterns we would expect if PIN is capturing the underlying information structure.

Given that the parameter estimates are stable across years, we pool the years to further illustrate the distribution of the parameters across stocks. Figure 4 shows these pooled distributions for our estimated parameters, and Table 1 presents summary statistics. It is clear from the figure that the composite parameter PIN is rather tightly distributed around the mode 0.18, while $\alpha$ and, in particular, $\delta$, are more dispersed over the parameter space. The skewness of $\delta$ is consistent with the generally rising stock prices over this period; since stocks typically did well, the probability of bad news was generally lower than that of good news. We have aggregated the uninformed trading variables to depict the balance between uninformed buying and selling. Over our time interval, uninformed traders were marginally more likely to sell, while informed traders were more likely to buy. This, too, is consistent with the economic conditions of our sample, as informed traders were better able to capture the benefits of good news and thus rising stock prices.

In summary, we have been able to estimate our structural model for a cross-section of stocks. The individual parameter estimates appear economically reasonable, and the small standard errors of our estimates indicate strong statistical significance. The time-series of our estimates indicate a remarkable stability, with very little year-to-year movement in our estimated parameters. Our contention is that the estimated variables measure the components of information-based trading, and their combination into our PIN variable provides a concrete measure of this risk for each stock.

A natural concern is that, while seemingly reasonable, these estimates are, by definition, unverifiable: information-based trading is not observable, and so our estimates could be artifacts of our estimating procedure, and not, as we claim, proxies for information. As with any model, however, the proof lies in its predictive power. In particular, if our estimates are measuring asymmetric information, then one obvious test is to see how well they do in explaining a phenomena known to be related to information: spreads. Equation (6) gives the predicted relationship between our estimated variables and opening spreads, and so an important evaluation of the model is how well it does in explaining actual spread behavior. Note that since our estimating procedure uses only trade data, and not prices, spreads provide an independent check on the validity of our approach.

### *4.3 Opening Spreads and Information-based Trading*

We collected opening bid and ask quotes from the ISSM and TAQ data bases for each stock in our sample for the time period 1983-1998. The data were filtered to exclude any likely errors. The percentage opening spread was then calculated as the ask less the bid quote divided by the quote midpoint. The daily distribution of the percentage opening spreads is given in the upper panel of Figure 5. The data vividly illustrate the impact of the market crash of October 19, 1987. While the immediate impact on spreads on that date is striking, a more intriguing finding is that spreads in the upper quartile widen and do not return to pre-1987 levels for many years. This contrasts sharply with median spreads and spreads in the lowest 5% of the distribution (which are typically those of the most active stocks), which quickly return to their pre-crash levels, and then actually show a slight downward trend. Consequently, the cross sectional distribution of spreads became more dispersed in the period following the 1987 crash.

The across stock average opening spreads also follow a different time series process after October 1987. First, there is an increased time series variation after the crash, as shown in the bottom panel of Figure 5, which depicts changes in the daily mean opening spread. The increase in variance is highly significant, as evidenced by the statistics from the variance homogeneity test given in Table 2. However, there is a more fundamental shift in the time series pattern of spread changes. Table 2 shows the standard deviation, skewness and excess kurtosis of the changes in the daily across stock mean spread. Not only does the standard deviation increase, but skewness and kurtosis are also much larger after October 1987, so that while the Shapiro-Wilk test does not reject that the data follow a normal distribution before 1987, it strongly rejects the null in the post-1987 period. In particular, the post-1987 data is skewed to the right, indicating that the upper tail is heavier than the lower tail, consistent with specialists now widening the spread very fast in response to perceived uncertainty, whereas when spreads are lowered, it is done in smaller steps.

Our model suggests two simple approaches for verifying how well our estimated parameters relate to actual spreads.[11] First, we can informally compare the pattern of actual spreads in Figure 5 with that of our predicted spread, as given by equation (6). Collating the definition of the percentage spread on the left hand side of equation (6) with our operational definition above, we note that the unconditional expected asset value, $V^*_i$, is proxied by the quote midpoint. The predicted spread on the right hand side of equation (6) depends on our estimated parameters and on the standard deviation of percentage returns. Thus, we calculated year by year for each stock in the sample the standard deviation of returns using the daily returns from the Center for Research in Security Prices (CRSP) daily files. We then computed our predicted percentage spread, which we denote by PISTD. Hence, we obtain one estimate of PISTD each calendar year for each stock in the sample. Figure 6 gives the yearly distribution of this variable. What is immediately striking is the similarity between the two series. Both predicted and actual spreads appear to jump in 1987, and while the medians recover, the upper quartiles widen. Furthermore, the magnitudes of the percentiles in the two figures are broadly similar, though the

[11] In previous research (EKOP [1996]) we examined the relationship between opening spreads and PIN for a small sample of stocks for a single year. Here we examine the time series and cross sectional relationship between spreads and PINs for nearly all NYSE stocks.

predicted spread does not attain the same spikes as that of the actual opening spread, which is not surprising as the predicted spread essentially is an average over all days of the year while the actual spread is shown on a daily basis. Since our model allows only information to influence spreads, we interpret these results as strong evidence in support of the economic reasonableness of our estimates.

A second approach to test our estimates is to use regression analysis. If our estimates actually reflect information-based trading, then they should be able to predict spreads. We ran the cross-sectional regression

$$(9) \qquad SPREAD_i = \beta_0 + \beta_1 PISTD_i + \nu_i$$

where $SPREAD_i$ is the mean of stock i's opening percentage spread over all trading days of the year, and $PISTD_i$ is the predicted percentage spread for stock i in that year. Table 3 - Panel A lists the regression parameter estimates for each year of our sample period.

The strong performance of the PISTD regressions provides compelling evidence of the reasonableness of our model and our estimates. The estimate of $\beta_1$ is positive and statistically significant, corresponding to our prediction that greater values of PISTD lead to higher spreads.. Moreover, the $R^2$ of the regressions are quite high, ranging from .41 to .71. A perfect fit for our model is $\beta_0 = 0$ and $\beta_1 = 1$, but as we find a positive intercept and a slope less than one, the predicted values are close but not exact. A plausible explanation for the positive intercept is simply that factors other than information affect opening spreads: inventory, specialist market power, and price discreteness are all likely culprits. The under-estimate of $\beta_1$ may reflect the econometric difficulties introduced by the regressor being stochastic. In particular, this problem produces a negative correlation between the true regressor and the error term, causing $\beta_1$ to be biased downward (and $\beta_0$ to be biased upward).

It is well known that spreads are also influenced by factors such as volume. Are we merely picking up volume effects with our PISTD variable, and not the information effects we claim? To address this concern, we ran the estimating equation

$$(10) \qquad SPREAD_i = \beta_0 + \beta_2 \, LOGVOL_i + \nu_i$$

where $LOGVOL_i$ is the logarithm of the average daily dollar trading volume for stock i in that year. The results are given in Table 3- Panel B. We find that volume enters with the expected effect, with the coefficient on $\beta_2$ being both negative and statistically significant. However, the $R^2$ of the LOGVOL regressions now range only from .38 to .46, significantly lower than the $R^2$ obtained when we use PISTD to predict spreads. Running the composite regression with both PISTD and LOGVOL (not reported) reveals that both variables retain their statistical significance, suggesting that both variables influence opening spreads.

These findings suggest that our model does a very good job of explaining actual spreads. For our perspective here, these results also provide strong confirming evidence linking our estimates with the underlying probability of information-based trading. Having established this link, we now turn to the deeper question of whether there is also a link between asset returns and information-based trading.

## 5. Asset Pricing Tests

### *5.1. Data and Methodology*

For the asset pricing tests, we need to use additional data on firm characteristics and returns. These data are available from the monthly CRSP files and the annual COMPUSTAT files. Data are not available for all of our listed firms, so the sample used in our asset pricing tests is drawn from the intersection of the NYSE listed firms on the CRSP and COMPUSTAT files. The monthly samples contain between 997 and 1316 stocks for the period 1984 to 1998, yielding 180 monthly observations to aggregate over time. One concern we note at the outset is the length of our sample period. Asset pricing tests typically employ long sample periods, but transaction data, which we need to calculate our PIN variable, are not available prior to 1983, and since we employ lagged PIN estimates, the asset pricing tests begin in 1984. Longer sample periods enhance the ability to find statistically significant factors influencing returns, so our limited sample period imposes a particularly stringent constraint on our testing approach.

To allow for comparability with previous work, our methodology follows that of Fama and French (1992) (FF). Fama and French explored the determinants of the cross-sectional variation in returns and found that beta, size, and book-to-market (i.e. the ratio of the book value of equity to the market value of equity) all influenced returns. Consequently, we include these variables, as well as our estimated PIN variable, in our analysis of asset pricing returns. We also explore whether the effect of PIN can be captured by previously suggested proxies for liquidity, namely bid-ask spreads and share turnover, or by return variation, by including these variables in the asset pricing regressions.

We calculate betas using the following approach. *Pre-ranking betas* are estimated for individual stocks using monthly returns from at least two years to, when possible, five years, before the test year. Thus, for each stock we use at least 24 monthly return observations in the estimation. We regress these stock returns on the contemporaneous and lagged value-weighted CRSP NYSE/Amex index. Pre-ranking betas are then given as the sum of the two coefficients (this approach, suggested by Dimson (1979), is intended to correct for biases arising from non-synchronous trading). Next, 20 portfolios are sorted every January on the basis of the estimated betas, and monthly portfolio returns are calculated as equal-weighted averages of individual stock returns. *Post-ranking portfolio betas* are estimated from the full sample period, such that one beta estimate is obtained for each of the 20 portfolios. Portfolio returns are regressed on contemporaneous and lagged values of CRSP index returns. The portfolio beta, $\hat{\beta}_p$, is then the sum of the two coefficients. We use individual stocks in the cross-sectional regressions, so individual stock betas are taken as the beta of the portfolio to which they belong. Because the portfolio compositions change each year, individual stock betas vary over time.

We calculate the other variables in our asset pricing tests as follows. Returns for each stock are taken from the CRSP monthly return files, using the CRSP de-listing return in the month of possible de-listing. All returns are in excess of the one-month T-bill rates. $PIN_{it-1}$ is the probability of information-based trading in year *t*-1 whose estimation is described in the previous section.[12] $SIZE_{it-1}$ is the logarithm of market value of equity in firm i at the end of year *t*-1. Book value of common equity is obtained from the annual COMPUSTAT files (item 60). Following Fama-French, we exclude firms with negative book values, and we set BE/ME values outside the 0.005 and .995 fractiles equal to these fractiles, respectively. We take logs, such that the explanatory variable, $BM_{it-1}$, is $\ln(BE_{t-1}/ME_{t-1})$ for firm *i*.

For each month in the sample period 1983-1998, we ran the following cross-sectional regression:

$$(11) \qquad R_{it} = \gamma_{0t} + \gamma_{1t}\hat{\beta}_p + \gamma_{2t}PIN_{it-1} + \gamma_{3t}SIZE_{it-1} + \gamma_{4t}BM_{it-1} + \eta_{it},$$

where $R_{it}$ is the excess return of stock i in month *l* of year *t* (monthly subscripts omitted), $\gamma_{jt}$, $j = 1,\ldots, 5$, are the estimated coefficients, and $\eta_{it}$ is the mean-zero error term. The coefficients from the cross-sectional regressions are averaged through time, using the standard Fama-MacBeth (1973) methodology. Because this procedure is inefficient under time-varying volatility, we also use the correction technique suggested by Litzenberger and Ramaswamy (1979). This correction weights the coefficients by their precisions when summing across the cross-sectional regressions, and is essentially a weighted least-square methodology. We report both the unadjusted and the Litzenberger-Ramaswamy adjusted coefficients.

A problem with almost all variables provided as alternatives to beta as the explanatory variable of the cross-section of returns (for example, size, book-to-market, earnings-to-price, turnover, etc.) is that these variables depend on the security price. Miller and Scholes [1982] noted that the inverse of price may be a good measure of the conditional beta, and therefore regression analysis may be capturing mis-measurement of beta, rather than some alternative priced factor. Berk [1995] makes a related point. Because the estimation of PIN involves only trades, we avoid this potential critique in our inclusion of PIN.

Our primary interest lies in the time series average of $\gamma_{2t}$, namely the coefficient for PIN. Our hypothesis is that a higher risk of information-based trading for a stock translates into a higher required return for that stock, so we expect a significantly positive average coefficient on PIN.

### *5.2. Results*

Summary statistics on the variables in the asset pricing regressions are provided in Table 4. The procedure on beta sorting portfolios resulted in a reasonably broad variation in beta, with beta ranging from between 0.57 and 1.32, and our mean value of 1.0 is as expected. As noted in the previous section, our estimated PIN variable has a mean of 0.19, while ranging from 0 to

[12] We use PIN rather than the more complex variables in equation (6) as it is independent of prices, unlike PISTD, and as it has an obvious interpretation as the probability of information-based trading.

0.53. The means of the Size and Book to Market variables are also consistent with prior work on this sample period.

We first investigate the inter-relationships of the explanatory variables, and in particular how PIN correlates with each variable. Table 5 present time series means of the monthly bivariate correlations of the variables in the asset pricing tests. One of the largest absolute correlations is between size and PIN, with an average correlation of –0.575. This finding confirms results from earlier research that larger firms tend to have lower probabilities of informed trading.[13] One might expect that stocks with greater private information have higher systematic volatility, and this appears to be the case: PIN is positively correlated with beta, with a correlation of 0.156. We had weaker priors on the relation between PIN and BM, but note a positive correlation (0.168). The correlation between return and PIN is rather low, but the correlation between return and the other explanatory variables is similarly low.

Return and beta are negatively correlated, but, as discussed below, this is in line with prior findings in this sample period. Likewise, the positive correlation between return and size is opposite of that reported in earlier periods, but it is consistent with findings from our sample period. Finally, the low correlation between return and BM is not unexpected given Loughran's [1997] finding that book-to-market arises primarily in Nasdaq stocks, and our sample uses only NYSE firms.

The results from the asset pricing tests are provided in Table 6. The results give striking evidence that that the risk of informed trading as captured by PIN is priced in the required returns of stocks. Looking at the weighted least squares results, we find a positive and significant coefficient on PIN (t-value 4.43). The point estimate of the PIN coefficient has the natural interpretation that a difference of 10 percentage points in PIN between two stocks translates into a difference in required return of 0.21 percent per month. This is an economically meaningful and significant difference. We also find a significant and positive coefficient on SIZE (t-value 9.57), and a significant, but negative, coefficient on BETA (t-value –5.87). This latter finding, while inconsistent with standard asset-pricing theory, is consistent with the findings of Fama and French (1992), Chalmers and Kadlec (1998) and Datar, Naik and Radcliffe (1998) who investigate similar sample periods. Book-to-Market is not significant, a finding not unexpected given our earlier discussion.

---

[13] See Easley, Kiefer, O'Hara, and Paperman [1996].

That PIN affects asset returns is consistent with the economic analysis motivating our work. We believe that the PIN variable captures aspects of the dynamic efficiency of stock prices. These dynamic effects arise because information-based trading affects not only the spread, but the evolution of prices as well. Our results are consistent with this dynamic efficiency influencing the required returns for stocks.

### *5.3 Alternative Explanations*

It is natural to ask whether PIN works in our asset pricing regressions because it is a fundamental priced variable, or because it is serving as a proxy for some omitted variable. There are three obvious candidates, and innumerable less obvious candidates, for the omitted variable designation. The most obvious candidate is spreads. We have shown that the probability of information-based trade is an important determinant of spreads. Earlier researchers (for example, Amihud and Mendelson (1986)) found a positive relation between spread and returns, so it could be that PIN is serving as a proxy for spread. Second, a stock with a high PIN is one with substantial imbalances in trades, and thus is a stock whose price is likely to be highly variable.[14] So it could be that PIN is serving as a proxy for the variability of returns on the stock. Of course, to the extent that this risk can be diversified away it should not be priced, and any nondiversifiable component of the risk should be picked up by $\beta$. But we know that CAPM does not work well over this time period, so this risk could be positively related to observed average excess returns. Finally, there has been substantial interest in the role of volume, or turnover, in explaining asset price behavior. Is PIN merely serving as a proxy for these measures? Earlier we found that PIN and volume each played a role in explaining spread behavior. We now consider whether these separate effects also hold in asset pricing.

We consider each of these variables in turn, and while we show that they are correlated with PIN as expected, we also show that when they are included in the returns regression they do not eliminate the direct effect of PIN. Specifically, we ran three regressions of the form

$$R_{it} = \gamma_{0t} + \gamma_{1t}\hat{\beta}_p + \gamma_{2t}PIN_{it-1} + \gamma_{3t}SIZE_{it-1} + \gamma_{4t}BM_{it-1} + \gamma_{5t}X_{it-1} + \eta_{it} \qquad (12)$$

[14] This relationship is not completely straightforward because although trade moves prices, public information events, which in our model do not generate trade, also move prices. So there could be stocks with low PIN, that is

where $R_{it}$ is the excess return of stock i in month *l* of year *t* (monthly subscripts omitted), $\gamma_{jt}$ are the estimated coefficients, $X_{it-1}$ is, respectively, the average opening percentage spread for stock i in year *t*-1; the standard deviation of daily returns for stock i in year *t*-1; or a turnover and volatility of turnover measure for stock i in year *t*-1; and $\eta_{it}$ is the mean-zero error term. Results of the regressions are reported in Table 7.

We first consider spread. Consistent with our analysis in Section 4, we define the variable $SPREAD_{it-1}$ to be the average of the daily opening percentage spreads for stock *i* in year *t-1*. We know from Table 5 that PIN and SPREAD are positively, but not perfectly, correlated. Indeed, the relatively low .353 correlation reflects that spreads can be affected by many factors other than information. Is PIN or SPREAD the better predictor of returns? We test this by first including SPREAD in place of PIN in the asset pricing regression and then by including both SPREAD and PIN in the regression. When SPREAD is included, and PIN is excluded, SPREAD is marginally significant, but its coefficient is negative.[15] This result is not what would be expected from liquidity-based theories of asset returns. When SPREAD and PIN are both included, SPREAD becomes insignificant and its coefficient remains negative. But PIN remains highly significant (t-value=4.0) with the correct positive sign. The inclusion of spread reduces the coefficient on PIN only slightly from 2.1 to 1.9.

That it is PIN, and not SPREAD, that affects asset pricing returns is consistent with the economic analysis motivating our work. While traders undoubtedly care about spreads, they are more concerned with the risk of holding the stock, and this is affected by the extent of private information. Information-based trading does give rise to spreads, but spreads in actual markets can be affected by many other factors such as minimum tick sizes, specialist continuity rules, and even market power. These factors dictate that spreads will be, at best, a noisy proxy for the risk of informed trading. We believe our results here provide strong evidence that information plays a deeper role, one beyond that captured, however imperfectly, in spreads.

We next consider whether PIN or the variability of returns on a stock is the better predictor of excess returns. We define $STD_{it-1}$ to be the standard deviation of daily returns on

---

little information based trade, and highly variable prices caused by the release of public information. This is consistent with our analysis in that public events can move the range of true values of the stock.

[15] This result is consistent with the findings of Chalmers and Kadlec (1998).

stock i in year *t*-1. As expected, Table 5 shows that STD and PIN are positively correlated, with a correlation coefficient of 0.240. When STD is included in the pricing regression in place of PIN, it is highly significant (TSTAT=-12.2) with a negative coefficient. This also changes the coefficient on $\beta$ from negative and significant to positive and insignificant. This occurs in part because the $\beta$ used in our regressions is the portfolio $\beta$ and not the individual stock's $\beta$. Of more importance for us is that when both STD and PIN are included the coefficient on PIN remains positive and significant (TSTAT=3.05). The effect of STD indicates the weakness of CAPM, or at least our standard implementation of it, over this period. But it has little effect on the pricing of the probability of information based trade.

Finally, we consider whether PIN can be interpreted as a proxy for volume effects. Volume can be measured in many ways, but previous research on asset pricing effects has typically used turnover, or daily volume divided by shares outstanding. This measure avoids any of the price-beta concerns noted earlier, and also allows for greater comparability across stocks. Datar, Naik and Radcliffe [1998] present evidence that there is a negative relationship between turnover and returns. Further, Anshuman, Chordia, and Subrahmanyam [2000] (ANS) argue that volume effects are better captured by allowing both turnover and the volatility of turnover to affect price behavior, and find that both variables negatively affect returns. We calculate share turnover in each stock for each month in year *t*-3 to *t*-1, as the number of shares traded divided by the number of shares outstanding. The natural logarithm of the average turnover is then used in the asset pricing regressions for year *t*. As a proxy for the variability of turnover, we follow ANS [2000] and employ the natural logarithm of the coefficient of variation of the monthly turnover in years *t*-3 to *t*-1. Including Turnover and CV turnover for each stock in the estimating equation (without PIN) reveals a strong negative effect on returns of both variables, similar to the findings in ANS [2000]. When PIN is included along with these variables, the coefficient on PIN remains positive and significant (t-stat 2.917).[16] Thus, it appears that the influence of PIN on returns is not proxying for the effects of volume.

In summary, the positive relationship between expected return and the probability of informed trading seems be to robust to the inclusion of different explanatory variables in the

[16] Including only turnover we find similar effects. Turnover alone enters negatively and significantly, while Turnover and PIN together are both significant (t-stats of -9.88 and 2.423, respectively) and retain their predicted signs.

cross-sectional regressions. Thus, there is evidence that the risk of informed trading is, indeed, an important determinant of the required stock returns.

## 6. Conclusions

We have investigated the role of information-based trading in affecting asset returns. Our premise is that in a dynamic market, asset prices are continually adjusting to new information. This evolution dictates that the process by which asset prices become informationally efficient cannot be separated from the process generating asset returns. Our theoretical model suggests that private information influences this price evolution, and in so doing affects the risk of holding the asset. We set out to test this link between asset prices and private information by using the structure of a sequential trade market microstructure model to derive an explicit measure of the probability of information-based trading for an individual stock. We then estimated this probability for a large sample of NYSE-listed stocks. Incorporating our probability estimates into a standard asset pricing framework revealed strong support for our hypothesis: Information-based trading has a large and significantly positive effect on asset returns. Indeed, our estimated information variable and firm size are the predominant factors explaining returns.

That the risk of information-based trading affects asset returns raises a host of important questions regarding asset pricing in general, and asset pricing models in particular. Brevity precludes addressing all of these, but we do feel it useful to consider three general issues. These involve the theoretical basis for our result, the empirical properties of PIN, and the implications of our results for future research.

Of particular importance is why this can occur in a seemingly efficient capital market. A natural objection to all candidates put forward to explain asset returns is that, with the exception of systematic risk, the actions of arbitrageurs should remove any such proposed influence on the market. While this may be accurate for some factors, it is not accurate with respect to asymmetric information. In a world with asymmetric information, an uninformed investor is always at a disadvantage relative to traders with better information. In bad times, this disadvantage can result in the uninformed trader's portfolio holding too much of the stock; in good times, the trader's portfolio has too little of the stock. Holding many stocks cannot remove this effect because the uninformed do not know the proper weights of each asset to hold. In this sense, asymmetric information risk is systematic because, like market risk, it cannot be diversified away.[17]

In our empirical work we found that PIN, the probability of information-based trading variable, actually dominated all other variables, including $\beta$, in explaining returns. Given our argument that information has a systematic component, this should not be unexpected. We caution, however, that our results do not mean that only private information matters in asset pricing. Our theoretical model demonstrates that this is not true; many factors affect the risk of holding assets, and so, too, should they affect asset pricing. What our results do suggest is that the effects of information may be more pervasive, and important, than our simple theories, and asset-pricing models, have thus far considered.

The success of our PIN variable naturally leads to questions regarding its empirical properties. A very useful exercise would be to examine the cross-sectional determinants of PIN, and in particular how PIN relates to variables such as industry or accounting measures. Not surprisingly, this is a large endeavor and one we hope to address in future work. One benefit of such a project could be to determine a set of "sufficient statistics" for PIN that involves accounting data. As is clear from this paper, the actual calculation of PIN requires a tremendous amount of computation. Replicating PIN with more easily available data would make it easier to apply, and would have the added benefit of explaining why it is that some accounting data appears to be informative for asset pricing.

Finally, our results here suggest a number of directions for future research. There is now a substantial body of work suggesting that volume, and volume-linked variables, play an important role in asset pricing. We have shown here that PIN is not a volume effect, but there remains the intriguing question of whether volume effects may be not be proxying for some of the underlying components of PIN such as the rate of uninformed trade or the probability of new information. Investigating the role of the components of PIN would provide insight into this issue. An equally intriguing issue is momentum. There is now wide-spread, if in some cases grudging, acceptance of the fact that momentum affects asset prices. These momentum

[17] It is also not the case in our model that the informed traders will simply trade the effect because they too face risk in holding the asset. Informed traders are also risk averse, and so there will always be a premium to hold the risky asset. However, because the stock is relatively less risky for the informed, in equilibrium their expected holdings of the asset exceed that of the uninformed. See Easley and O'Hara [2000] for more analysis and discussion.

effects appear to arise over relatively short time intervals (months 3-12), and they pose a challenge for virtually all asset pricing theories. One possible explanation is that momentum is somehow linked to the underlying information structure of the stock. Testing for such effects using our approach would require finer estimates (i.e. monthly) of our PIN variable, as well as potentially a longer time frame. We hope to consider this in future research.

## Appendix

In this section we provide the derivations of bid and ask prices that lead to the spread equation (6) in the text. Let P(t) = ($P_n$(t), $P_b$(t), $P_g$(t)) be the market maker's belief about the events "no information event" (n), "bad news" (b), and "good news" (g) conditional on the history of trade prior to time t. So P(0) = (1–α, αδ, α(1–δ)). The expected value of the asset on day i conditional on the history of trade prior to time *t* is thus

(A.1) $$E[V_i \mid t] = P_n(t)V_i^* + P_b(t)\underline{V}_i + P_g(t)\overline{V}_i.$$

At any time t, the zero expected profit bid price, b(t), is the market maker's expected value of the asset conditional on the history prior to *t* and on the arrival of an order to sell at t. Calculation shows that the bid at time *t* is,

(A.2) $$b(t) = E[V_i \mid t] - \frac{\mu P_b(t)}{\varepsilon_S + \mu P_b(t)}\left(E[V_i \mid t] - \underline{V}_i\right)$$

Similarly, the ask at time t, a(t), is the market maker's expected value of the asset conditional on the history prior to *t* and on the arrival of an order to buy at time t. Thus the ask at time *t* is

(A.3 $$a(t) = E[V_i \mid t] + \frac{\mu P_g(t)}{\varepsilon_B + \mu P_g(t)}\left(\overline{V}_i - E[V_i \mid t]\right).$$

These equations demonstrate the explicit role played by arrival rates for informed and uninformed traders in determining trading prices. If there are no informed traders (μ = 0), then trade carries no information, and so the bid and ask are both equal to the prior expected value of the asset. Alternatively, if there are no uninformed sellers ($\varepsilon_S$=0) then $b(t) = \underline{V}_i$ ; similarly, if the are no uninformed buyers then $a(t) = \overline{V}_i$ for all t. Generally, both informed and uninformed traders will be in the market, and so the bid is below $E[V_i \mid t]$ and the ask is above $E[V_i \mid t]$. This spread results from the market maker setting prices to protect her from expected losses to informed traders.

The factors influencing the spread are easier to identify if we write the spread explicitly. Let $\Sigma(t) = a(t) - b(t)$ be the spread at time t. Then

(A.4) $$\Sigma(t) = \frac{\mu P_g(t)}{\varepsilon_B + \mu P_g(t)}\left(\overline{V}_i - E[V_i \mid t]\right) + \frac{\mu P_b(t)}{\varepsilon_S + \mu P_b(t)}\left(E[V_i \mid t] - \underline{V}_i\right).$$

The spread at time *t* can be viewed in two parts. The first term is the probability that a buy is information - based times the expected loss to an informed buyer, and the second is a symmetric term for sells.

The percentage spread for the opening quotes will be important in our empirical work. It is

(A.5) $$\frac{\Sigma(0)}{V^*_i} = \mu\sqrt{\alpha\delta(1-\delta)}\left[\frac{1}{\varepsilon_B + \mu\alpha(1-\delta)} + \frac{1}{\varepsilon_S + \mu\alpha\delta}\right]\sigma_V$$

where $\sigma_V$ is the standard deviation of the daily percentage changes in the value of the asset. This standard deviation reflects the potential loss to the market maker from trading with informed traders. The remaining terms in the spread equation reflect the risk of trading with an informed trader. This risk is clearly a crucial factor influencing the size of spreads.

Finally, we note that when $\varepsilon_b = \varepsilon_s = \varepsilon$ and $\delta = 0.5$ equation A.5 simplifies to

(A.6) $$\frac{\Sigma(0)}{V^*_i} = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon}\left(\frac{\overline{V}_i - \underline{V}_i}{V_i^*}\right)$$

## Table 1: Parameter Summary Statistics

The table contains time series averages across years 1983-98 of cross sectional means, medians, standard deviations, and the median of parameter standard errors from the likelihood estimation.

| Variable | Mean | Median | StDev | Median StErr |
|---|---|---|---|---|
| alpha | 0.283 | 0.281 | 0.111 | 0.035 |
| delta | 0.331 | 0.309 | 0.181 | 0.066 |
| mu | 31.075 | 21.303 | 32.076 | 0.996 |
| eb | 22.304 | 11.437 | 31.519 | 0.324 |
| es | 24.046 | 13.095 | 31.427 | 0.299 |
| PIN | 0.191 | 0.185 | 0.057 | 0.019 |

## Table 2: Opening Spreads

The table contains statistics on the changes in the daily across stock mean spread. The sample period is 1983-98 with the month of October 1987 excluded, and statistics on the two subsamples before and after October 1987 are calculated. The Shapiro-Wilk tests that the data come from a normal distribution. The Brown-Forsythe Modified Levene is a test for equality of variance between the two subperiods, and it follows a $F_{(1,3866)}$ distribution.

| | Period | | |
|---|---|---|---|
| | Full | Pre Oct. 1987 | Post Oct. 1987 |
| N | 3868 | 1071 | 2797 |
| St. Dev. | 0.049 | 0.040 | 0.053 |
| Skewness | 0.499 | 0.149 | 0.548 |
| Excess Kurtosis | 6.027 | 1.095 | 6.181 |
| Shapiro-Wilk p-value | 0.001 | 0.971 | 0.001 |
| Brown-Forsythe modified Levene F-value | 43.08 | | |

## Table 3: Spread regressions

The table contains statistics from the cross-sectional regressions

$$SPREAD_i = \beta_0 + \beta_1 PISTD_i + \nu_i \text{, and}$$

$$SPREAD_i = \beta_0 + \beta_2 LOGVOL_i + \nu_i$$

where $SPREAD_i$ is the mean percentage opening spread for stock i during year t, $PISTD_i$ is the percentage opening spread predicted by the structural model, and defined by equation (6), and $LOGVOL_i$ is the logarithm of the average daily dollar trading volume for stock i in year t. Regressions are run each year in the sample period 1983-98 and with all years pooled. White heteroskedasticity consistent standard errors are given in parentheses.

| | Panel A | | | Panel B | | |
|---|---|---|---|---|---|---|
| Year | Intercept | Pistd | Adj. $R^2$ | Intercept | Logvol | Adj. $R^2$ |
| All | 0.29 | 0.83 | 0.598 | 8.95 | -0.52 | 0.406 |
| | ( 0.03) | ( 0.02) | | ( 0.11) | ( 0.01) | |
| 1983 | 0.43 | 0.68 | 0.409 | 6.01 | -0.34 | 0.383 |
| | ( 0.08) | ( 0.06) | | ( 0.24) | ( 0.02) | |
| 1984 | 0.53 | 0.70 | 0.455 | 7.05 | -0.41 | 0.409 |
| | ( 0.09) | ( 0.07) | | ( 0.29) | ( 0.02) | |
| 1985 | 0.31 | 0.86 | 0.643 | 7.46 | -0.44 | 0.430 |
| | ( 0.04) | ( 0.04) | | ( 0.30) | ( 0.02) | |
| 1986 | 0.01 | 0.93 | 0.584 | 8.58 | -0.50 | 0.395 |
| | ( 0.09) | ( 0.07) | | ( 0.42) | ( 0.03) | |
| 1987 | 0.14 | 0.68 | 0.565 | 8.43 | -0.48 | 0.417 |
| | ( 0.09) | ( 0.05) | | ( 0.35) | ( 0.02) | |
| 1988 | 0.46 | 0.82 | 0.487 | 9.63 | -0.57 | 0.438 |
| | ( 0.07) | ( 0.06) | | ( 0.42) | ( 0.03) | |
| 1989 | 0.44 | 0.80 | 0.576 | 9.52 | -0.56 | 0.419 |
| | ( 0.10) | ( 0.07) | | ( 0.46) | ( 0.03) | |
| 1990 | 0.25 | 0.93 | 0.714 | 11.59 | -0.70 | 0.455 |
| | ( 0.09) | ( 0.06) | | ( 0.53) | ( 0.04) | |
| 1991 | 0.30 | 0.93 | 0.727 | 12.19 | -0.73 | 0.456 |
| | ( 0.08) | ( 0.05) | | ( 0.51) | ( 0.03) | |
| 1992 | 0.53 | 0.79 | 0.586 | 12.02 | -0.72 | 0.445 |
| | ( 0.17) | ( 0.11) | | ( 0.59) | ( 0.04) | |
| 1993 | 0.13 | 0.99 | 0.668 | 9.93 | -0.58 | 0.438 |
| | ( 0.07) | ( 0.05) | | ( 0.48) | ( 0.03) | |
| 1994 | 0.23 | 0.96 | 0.612 | 8.90 | -0.51 | 0.437 |
| | ( 0.07) | ( 0.06) | | ( 0.37) | ( 0.02) | |
| 1995 | 0.49 | 0.72 | 0.533 | 8.49 | -0.49 | 0.422 |
| | ( 0.11) | ( 0.09) | | ( 0.36) | ( 0.02) | |
| 1996 | 0.22 | 0.85 | 0.603 | 8.32 | -0.48 | 0.397 |
| | ( 0.07) | ( 0.06) | | ( 0.40) | ( 0.03) | |
| 1997 | 0.21 | 0.77 | 0.627 | 7.82 | -0.45 | 0.406 |
| | ( 0.07) | ( 0.06) | | ( 0.38) | ( 0.02) | |
| 1998 | 0.07 | 0.81 | 0.613 | 7.85 | -0.45 | 0.420 |
| | ( 0.06) | ( 0.05) | | ( 0.35) | ( 0.02) | |

**Table 4: Summary statistics**

The table contains means, medians, minimum value and maximum values on the variables included in the asset pricing regressions given by equations (11) and (12). All statistics are calculated from the full sample, that is, pooling all months. Return is the percentage monthly return in excess of the one-month T-bill rate. Betas are portfolio betas estimated from the full period using 20 portfolios. Pin is the probability of informed trading given by equation (7). Size is the natural logarithm of year-end market value of equity, and BM is the natural logarithm of book value of equity divided by market value of equity and trimmed at the 0.005 and 0.995 fractiles. Spread is the yearly average of the daily opening spreads in each stock. Std is the daily return standard deviation for stock *i* in year *t*. Turnover is the natural logarithm of the average monthly turnover year *t*-3 to *t*-1, and CVturn is the coefficient of variation of the monthly turnover year *t*-3 to *t*-1.

| Variable | Mean | Median | Min | Max |
|---|---|---|---|---|
| Return | 0.74 | 0.47 | -100.60 | 339.69 |
| Beta | 1.00 | 1.03 | 0.57 | 1.32 |
| Pin | 0.19 | 0.18 | 0.00 | 0.53 |
| Spread | 1.52 | 1.14 | 0.14 | 15.07 |
| Size | 13.29 | 13.30 | 6.65 | 18.62 |
| Bm | -0.52 | -0.47 | -3.35 | 2.39 |
| Std | 2.10 | 1.88 | 0.46 | 14.92 |
| Turnover | 1.56 | 1.59 | -2.33 | 4.37 |
| CVturn | -0.68 | -0.69 | -2.07 | 1.30 |

**Table 5: Simple Correlations**

The table contains the time series means of monthly bivariate correlations of the variables in the asset pricing tests.

| | Beta | Pin | Size | BM | Spread | Std | Turn-over | CVturn |
|---|---|---|---|---|---|---|---|---|
| Return | -0.013 | -0.006 | 0.022 | -0.005 | -0.020 | -0.036 | -0.021 | -0.018 |
| Beta | | 0.157 | -0.188 | 0.006 | 0.200 | 0.406 | 0.283 | 0.068 |
| Pin | | | -0.576 | 0.168 | 0.353 | 0.240 | -0.187 | 0.412 |
| Size | | | | -0.384 | -0.708 | -0.494 | 0.123 | -0.545 |
| BM | | | | | 0.274 | 0.113 | -0.030 | 0.161 |
| Spread | | | | | | 0.748 | -0.116 | 0.396 |
| Std | | | | | | | 0.294 | 0.332 |
| Turnover | | | | | | | | -0.006 |

**Table 6: Asset pricing tests**

The table contains time series averages of the coefficients in cross-sectional asset pricing tests using standard Fama-MacBeth (1973) methodology and Litzenberger-Ramaswamy (1979) precision weighted means.

$$R_{it} = \gamma_{0t} + \gamma_{1t}\hat{\beta}_p + \gamma_{2t}PIN_{it-1} + \gamma_{3t}SIZE_{it-1} + \gamma_{4t}BM_{it-1} + \eta_{it},$$

where $R_{it}$ is the excess return of stock i in month *l* of year *t* (monthly subscripts omitted), $\gamma_{jt}$, j = 1,…, 5, are the estimated coefficients, and $\eta_{it}$ is the mean-zero error term. Betas are portfolio betas calculated from the full period using 20 portfolios. $PIN_{it-1}$ is the probability of information-based trading in stock *i* of year *t*-1. $SIZE_{it-1}$ is given as the logarithm of market value of equity in firm *i* at the end of year *t*-1, and $BM_{it-1}$ is ln($BE_{t-1}$/$ME_{t-1}$) for firm *i* in year *t*-1. T-values are given in parentheses.

| | Beta | Pin | Size | BM |
|---|---|---|---|---|
| Fama-MacBeth | -0.201 | 1.768 | 0.150 | 0.029 |
| | (-.413) | (2.473) | (2.593) | (0.270) |
| L-R WLS | -0.600 | 2.101 | 0.160 | 0.025 |
| | (-5.87) | (4.426) | (9.567) | (0.681) |

## Table 7: Alternative Explanations

The table contains time series averages of the coefficients in cross-sectional asset pricing tests using standard Fama-MacBeth (1973) methodology and Litzenberger-Ramaswamy (1979) precision weighted means.

$$R_{it} = \gamma_{0t} + \gamma_{1t}\hat{\beta}_p + \gamma_{2t}PIN_{it-1} + \gamma_{3t}SIZE_{it-1} + \gamma_{4t}BM_{it-1} + \gamma_{5t}X_{it-1} + \eta_{it},$$

where $R_{it}$ is the excess return of stock i in month *l* of year *t* (monthly subscripts omitted), $\gamma_{jt}$, j = 1,…, 4, and x, are the estimated coefficients, and $\eta_{it}$ is the mean-zero error term. Betas are portfolio betas calculated from the full period using 20 portfolios. $PIN_{it-1}$ is the probability of information-based trading in stock *i* of year *t*-1. $SIZE_{it-1}$ is given as the logarithm of market value of equity in firm *i* at the end of year *t*-1, and $BM_{it-1}$ is $\ln(BE_{t-1}/ME_{t-1})$ for firm *i* in year *t*-1. The control variable $X_{it-1}$ is, respectively, in panel A, the average opening percentage spread for firm *i* in year *t*-1, and in panel B, the standard deviation of daily returns for firm *i* in year *t*-1. In panel C, we include the variables Turnover and the Coefficient of variation of turnover (CVturn). Turnover is the natural logarithm of the average monthly turnover year *t*-3 to *t*-1, and CVturn is the coefficient of variation of the monthly turnover year *t*-3 to *t*-1. T-values are given in parentheses.

| Panel A. | Beta | Pin | Size | BM | Spread |
|---|---|---|---|---|---|
| Fama-MacBeth | -0.171 | | 0.105 | 0.013 | -0.043 |
| | (-.356) | | (2.134) | (0.125) | (-.541) |
| L-R WLS | -0.535 | | 0.082 | 0.016 | -0.036 |
| | (-5.23) | | (4.433) | (0.441) | (-1.62) |
| Fama-MacBeth | -0.199 | 1.694 | 0.144 | 0.020 | -0.034 |
| | (-.414) | (2.513) | (2.868) | (0.195) | (-.429) |
| L-R WLS | -0.559 | 1.909 | 0.125 | 0.023 | -0.027 |
| | (-5.44) | (3.999) | (5.876) | (0.628) | (-1.20) |

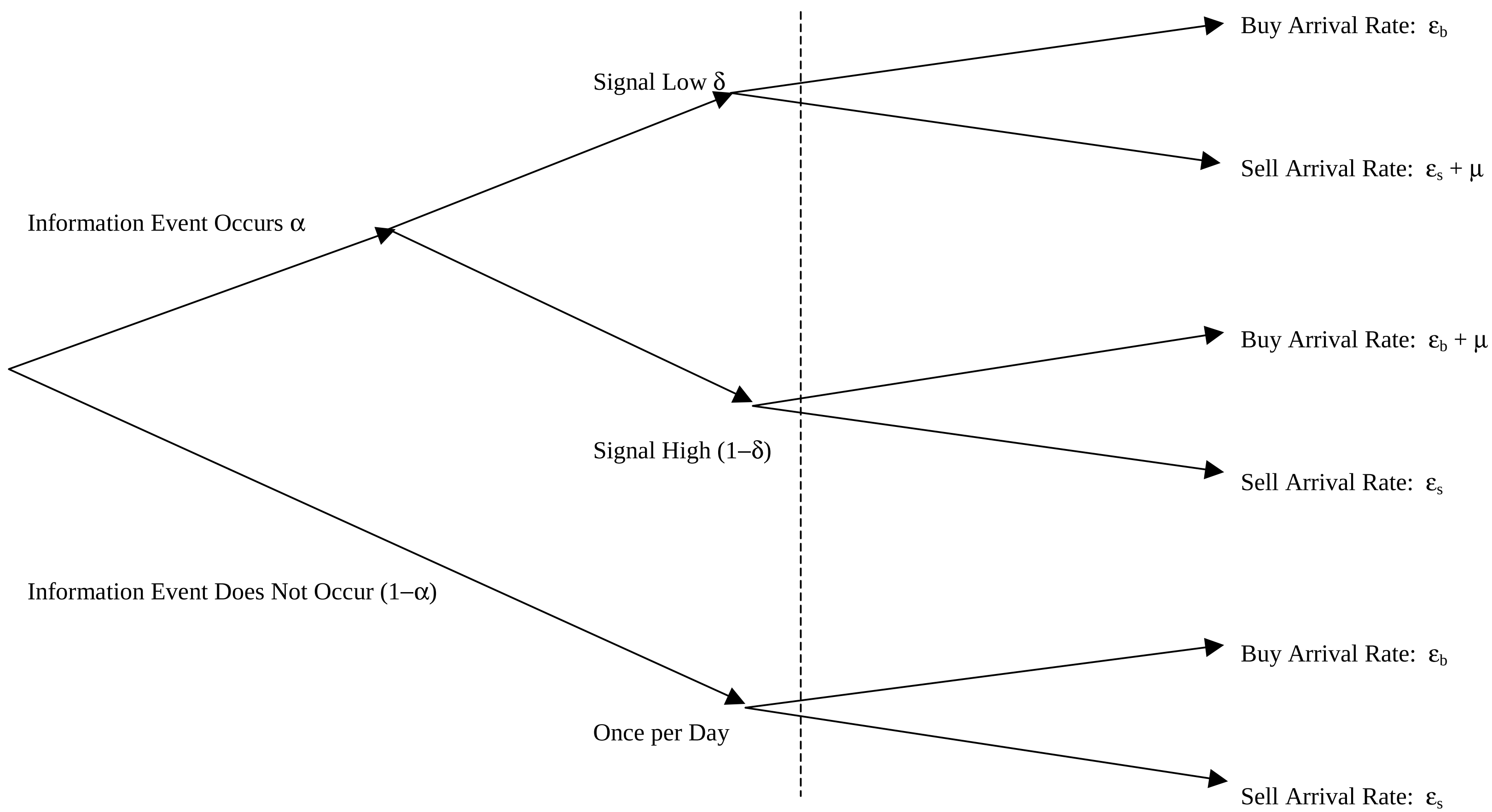
Panel B.

| | Beta | Pin | Size | BM | Std |
|---|---|---|---|---|---|
| Fama-MacBeth | 0.195 | | 0.047 | -0.033 | -0.301 |
| | (0.444) | | (1.014) | (-.321) | (-2.65) |
| L-R WLS | -0.010 | | 0.029 | -0.033 | -0.327 |
| | (-.089) | | (1.851) | (-.877) | (-12.2) |
| Fama-MacBeth | 0.168 | 1.409 | 0.077 | -0.026 | -0.294 |
| | (0.385) | (2.144) | (1.630) | (-.254) | (-2.59) |
| L-R WLS | -0.033 | 1.451 | 0.060 | -0.027 | -0.319 |
| | (-.302) | (3.050) | (3.270) | (-.715) | (-11.8) |

Panel C.

| | Beta | Pin | Size | BM | Turnover | CVTurn |
|---|---|---|---|---|---|---|
| Fama-MacBeth | -0.084 | | 0.108 | 0.028 | -0.314 | -0.211 |
| | (-.192) | | (1.946) | (0.268) | (-3.37) | (-2.51) |
| L-R WLS | -0.293 | | 0.125 | 0.028 | -0.344 | -0.178 |
| | (-2.71) | | (7.319) | (0.755) | (-10.3) | (-2.76) |
| Fama-MacBeth | -0.054 | 1.255 | 0.127 | 0.031 | -0.298 | -0.243 |
| | (-.126) | (1.645) | (2.201) | (0.300) | (-3.11) | (-2.91) |
| L-R WLS | -0.323 | 1.425 | 0.147 | 0.030 | -0.324 | -0.213 |
| | (-2.96) | (2.917) | (7.908) | (0.818) | (-9.53) | (-3.25) |

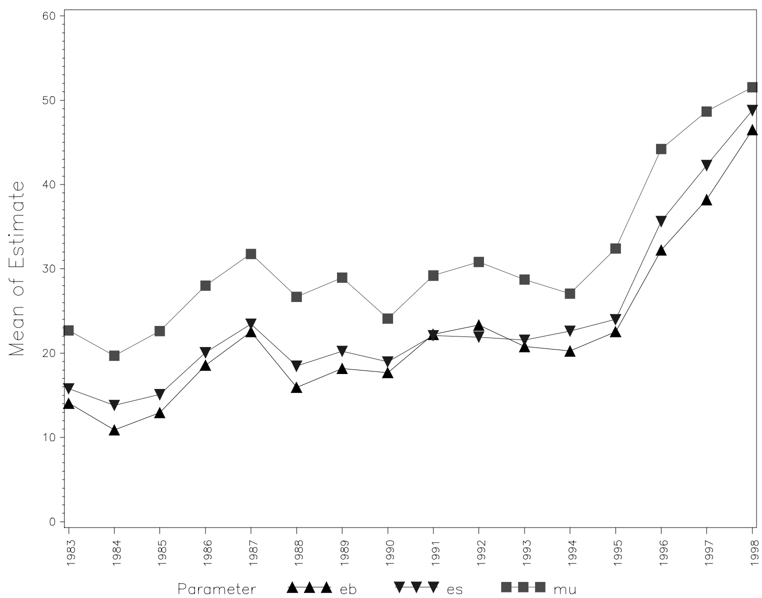
**Figure 1. Tree Diagram of the Trading Process.** $\alpha$ is the probability of an information event, $\delta$ is the probability of a low signal, $\mu$ is the rate of informed trade arrival, $\varepsilon_b$ is the arrival rate of uninformed buy orders and $\varepsilon_s$ is the arrival rate of uninformed sell orders. Nodes to the left of the dotted line occur once per day.



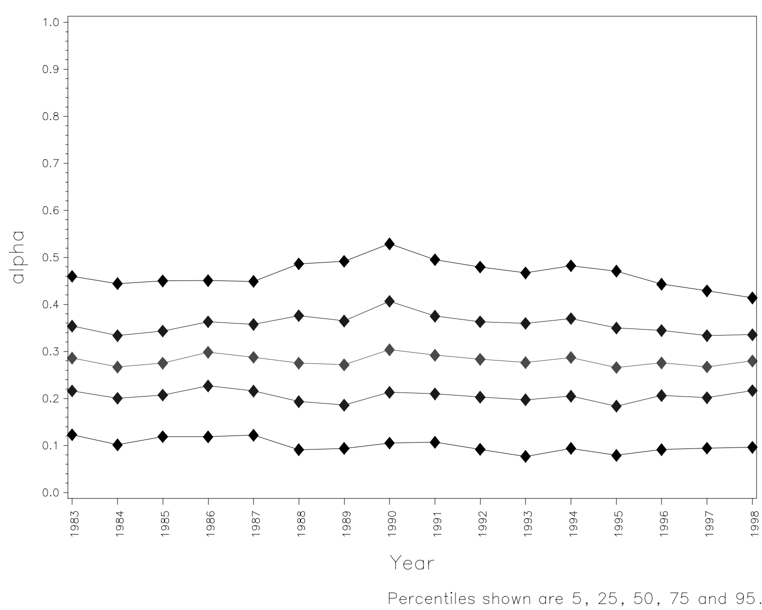
## Figure 2: Parameter distributions

The figure shows the cross-sectional distribution over time of the estimated parameters in the microstructure model given by the likelihood function in equation (2). Panel A gives the annual cross-sectional mean of the trading frequency parameters, $\varepsilon_b$, $\varepsilon_s$ and $\mu$. Panel B shows the 5th, 25th, 50th, 75th and 95th percentiles each year in the sample period for the cross-sectional distribution of $\alpha$, the probability that an information event has occurred. Panel C shows same percentiles for $\delta$, the probability of an information day containing bad news.

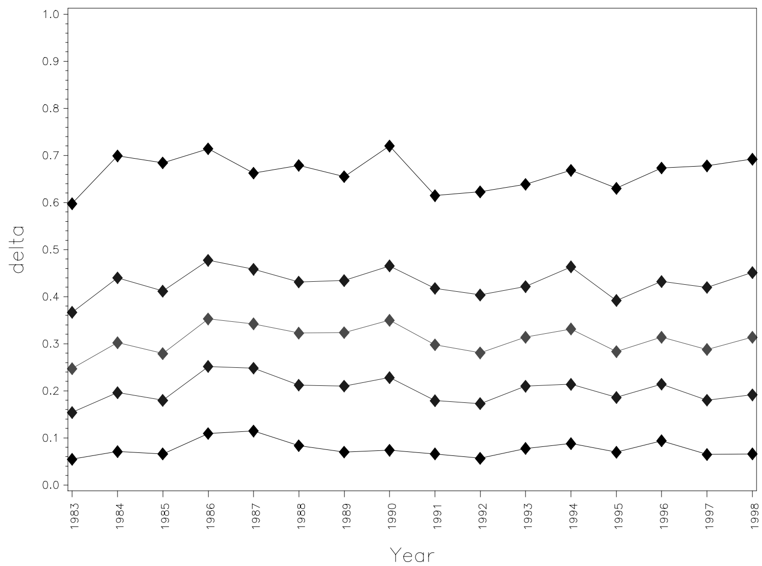
Panel A: Yearly means of trading frequency parameters, $\varepsilon_b$, $\varepsilon_s$ and $\mu$..




Panel B: Yearly distributions of the probability of information event, $\alpha$.



Panel C: Yearly distributions of the probability of an information day containing bad news, $\delta$.

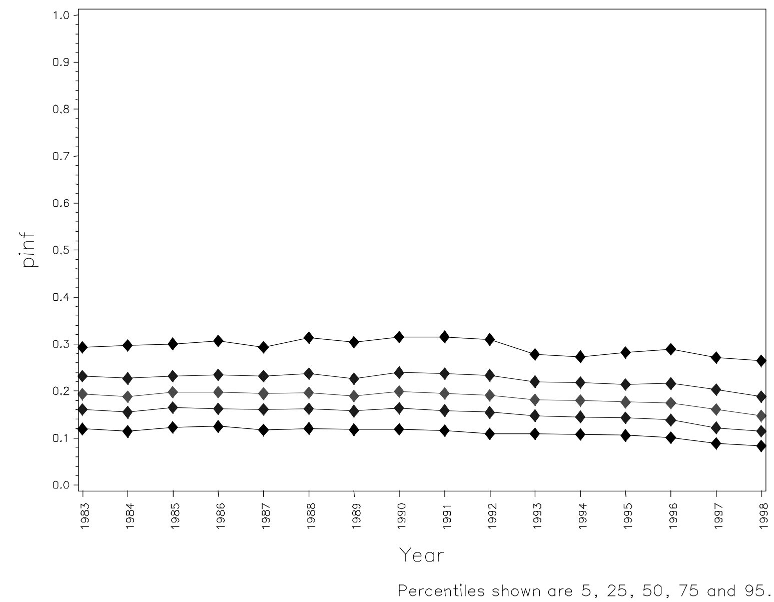


Percentiles shown are 5, 25, 50, 75 and 95.

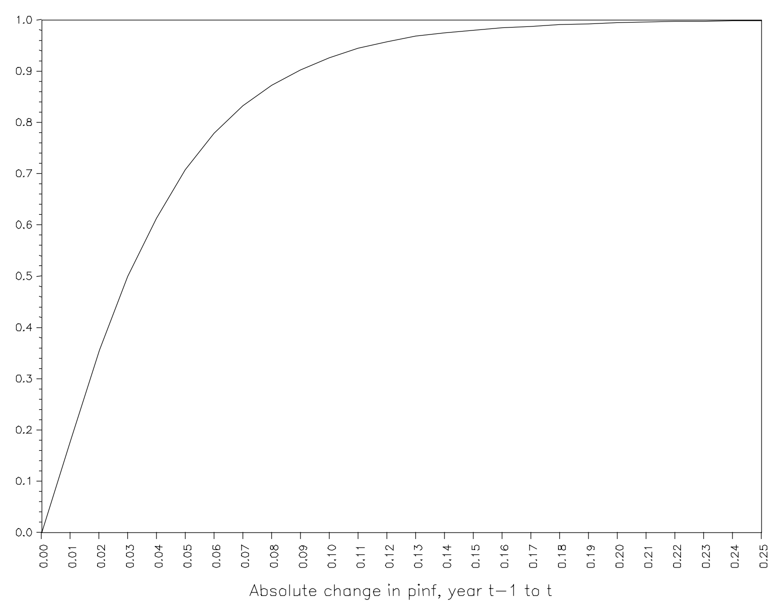
## Figure 3: PIN distributions

The figure shows the cross-sectional distribution of the estimated probability of information based trading, PIN, given by equation (7). Panel A shows the 5$^{th}$, 25$^{th}$, 50$^{th}$, 75$^{th}$ and 95$^{th}$ percentiles each year in the sample period for the cross-sectional distribution of PIN. Panel B shows the cumulative distribution of absolute price changes from year *t*-1 to year *t* of individual stock PIN estimates.

Panel A: Yearly distribution of PIN.



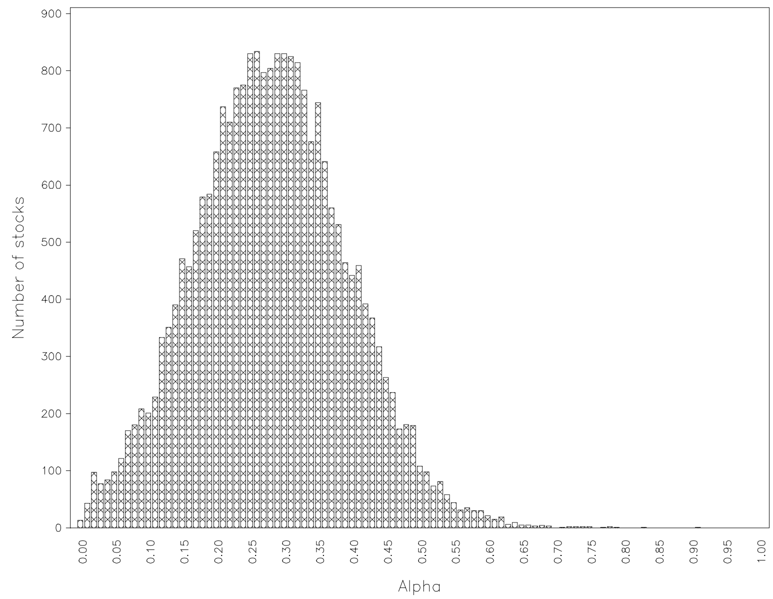

Panel B: Cumulative distribution of yearly absolute changes in PIN.



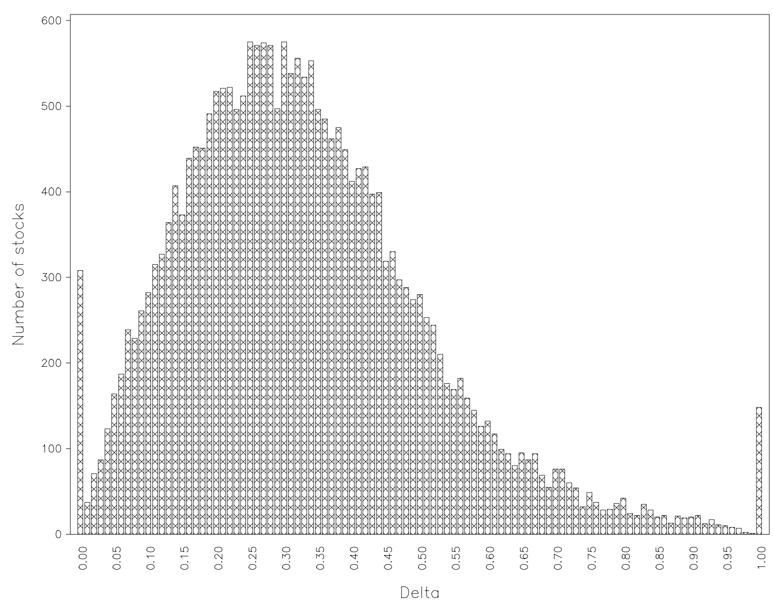
## Figure 4: Parameter distributions with pooled data

The figure shows the empirical distribution of the microstructure model parameters with all stocks and all years pooled. Panel A gives the distribution of $\alpha$, the probability that an information event has occurred. Panel B shows the distribution for $\delta$, the probability of an information day containing bad news. Panel C contains the distribution for PIN. Panel D shows the distribution of the uninformed order flow imbalance, $(\varepsilon_b\text{-}\varepsilon_s)/(\varepsilon_b+\varepsilon_s)$.

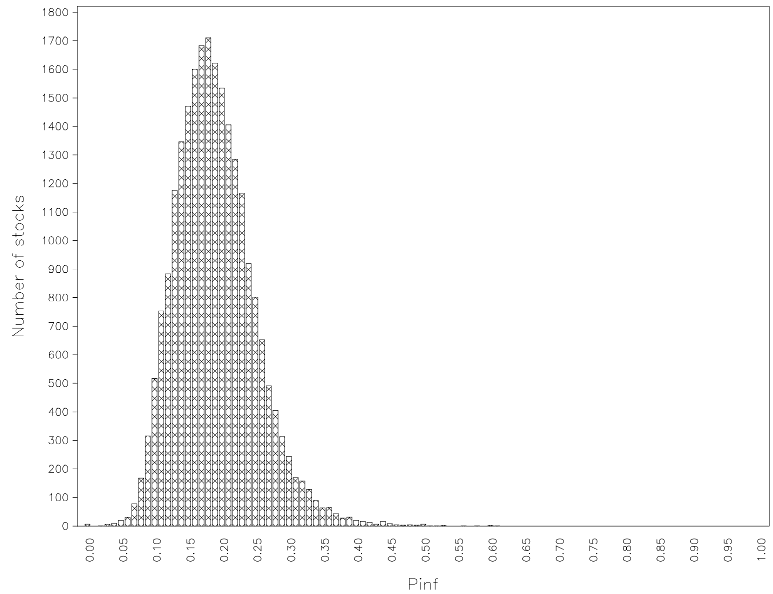
Panel A: Distribution of $\alpha$, the probability that an information event has occurred.



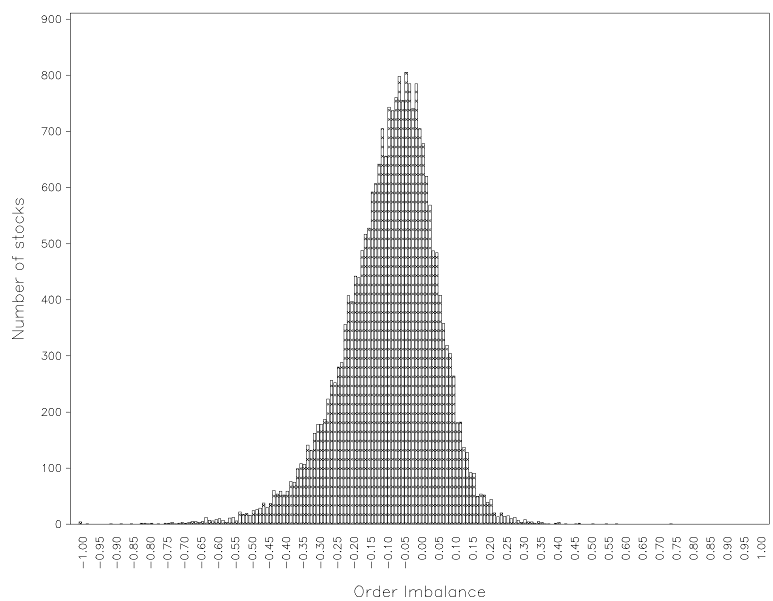

Panel B: Distribution of $\delta$, the probability of an information day containing bad news.



Panel C: Distribution of PIN.



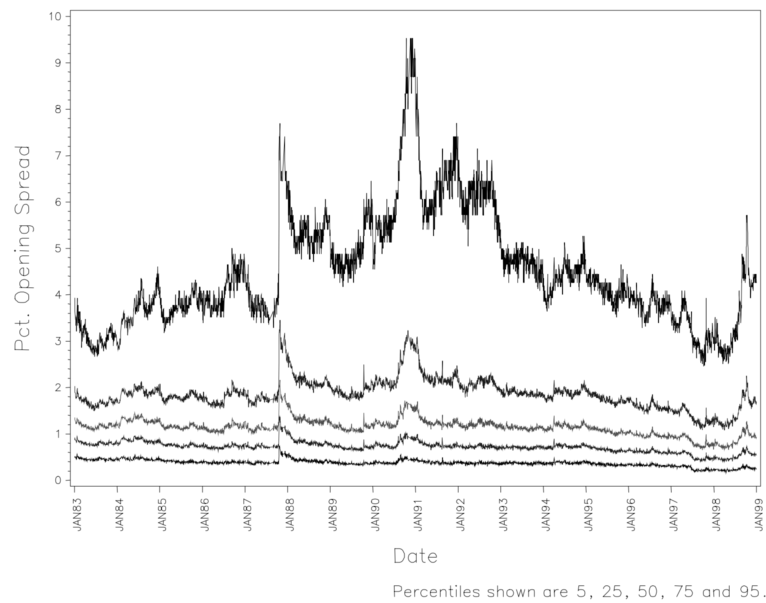

Panel D: Distribution of uninformed order flow imbalance, $(\varepsilon_b\text{-}\varepsilon_s)/(\varepsilon_b+\varepsilon_s)$.



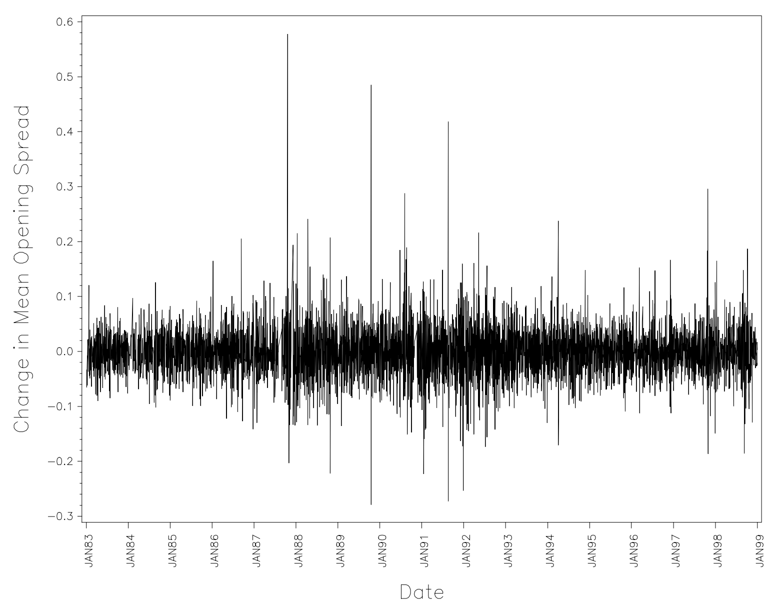
## Figure 5: Opening spreads

The figure shows opening percentage spreads in the sample of NYSE stocks retrieved from ISSM. Panel A contains the daily cross-sectional distribution represented by the 5th, 25th, 50th, 75th and 95th percentiles. Panel B indicates the volatility pattern by showing the first difference of the mean percentage opening spread.

Panel A: Daily cross-sectional distribution of opening spreads.

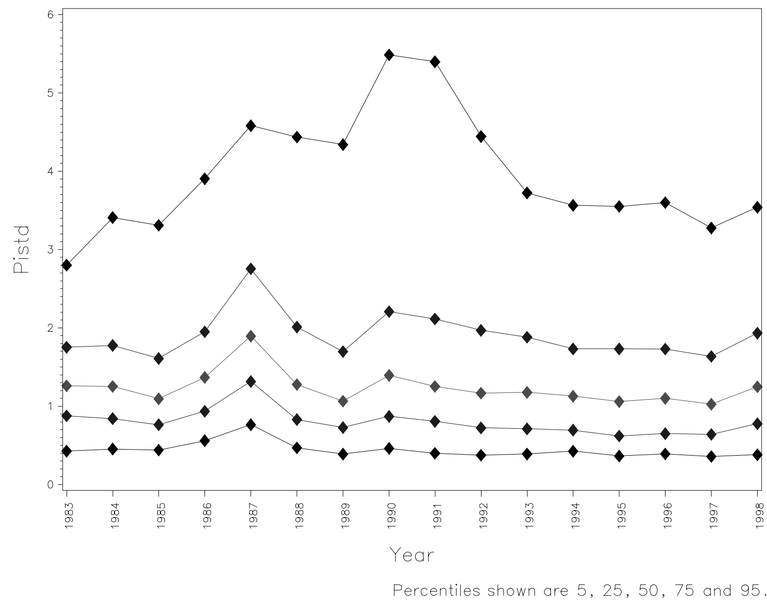



Panel B: Daily change in the mean percentage opening spread.



# Figure 6: Distribution of PISTD

The figure contains percentiles from the cross-sectional distribution of PISTD for each year in the sample period 1983-98. PISTD is defined by equation (5), and it is equal to the opening spread predicted by the structural model.



## References

Amihud, Y., 2000, "Illiquidity and Stock Returns: Cross-sectional and Time Series Effects," Working Paper, New York University.

Amihud, Y., and H. Mendelson, 1986, "Asset Pricing and the Bid-Ask Spread," *Journal of Financial Economics*, 17, 223-249.

Amihud, Y., and H. Mendelson, 1989, "The Effects of Beta, Bid-Ask Spread, Residual Risk, and Size on Stock Returns", *Journal of Finance*, 44(2), 479-486.

Amihud, Y., H. Mendelson, B. Lauterbach, 1997, "Market microstructure and securities values: Evidence from the Tel Aviv Stock Exchange," *Journal Of Financial Economics* 45(3), 365-390

Anshuman, V. R., T. Chordia and A. Subrahmanyam, 2000, "The Volatility of Trading Activity and Expected Stock Returns," Working Paper, Indian Institute of Management, Bangalore.

Berk, J. B., 1995. A critique of size-related anomalies, *Review of Financial Studies*, 8(2), 275-286

Brennan, M. J., T. Chordia, A. Subrahmanyam, 1998, "Alternative factor specifications, security characteristics, and the cross-section of expected stock returns," *Journal of Financial Economics*, 49(3), 345-373

Brennan, M. J., and A. Subrahmanyam, 1996, "Market microstructure and asset pricing: On the compensation for illiquidity in stock returns," *Journal of Financial Economics*, 41, 441-464.

Chalmers, J. M., and G. B. Kadlec, 1998, "An empirical examination of the amortized spread," *Journal of Financial Economics*, 48, 159-188.

Chen, N. F., and R. Kan, 1996, “Expected Return and the Bid-Ask Spread,” in *Modern Portfolio Theory and Applications*, ed. by K. S. S. Saitou, and K. Kubota. Gakujutsu Shuppan Center, Osaka.

Datar, V., N. Naik, and R. Radcliffe, 1998, “Liquidity and stock returns: An alternative test,” *Journal of Financial Markets*, 1(2), 203-219.

Dimson, E., 1979, “Risk Measurement When Shares Are Subject to Infrequent Trading,” *Journal of Financial Economics*, 7(2), 197-226

Easley, D., N. M. Kiefer, and M. O'Hara, 1996, “Cream-Skimming or Profit-Sharing? The Curios Role of Purchased Order Flow,” *Journal of Finance*, 51(3), 811-833.

——, 1997a, “The information content of the trading process,” *Journal of Empirical Finance*, 4, 159-186.

——, 1997b, “One Day in the Life of a Very Common Stock,” *Review of Financial Studies*, 10(3), 805-835.

Easley, D., N. M. Kiefer, M. O'Hara, and J. Paperman, 1996, “Liquidity, Information, and Less-Frequently Traded Stocks” *Journal of Finance*, 51, 1405-1436.

Easley, D., M. O'Hara, and J. Paperman, 1998, “Financial Analysts and Information-based Trade”, *Journal of Financial Markets*, 1(2), 175-201

Easley, D., and M. O'Hara, 2000, "Information and the Cost of Capital," Working Paper, Johnson Graduate School of Management, Cornell University.

Eleswarapu, V. R., 1997, “Cost of Transacting and Expected Returns in the Nasdaq Market,” *Journal of Finance*, 52(5), 2113-2127.

Eleswarapu, V. R., and M. R. Reinganum, 1993, “The seasonal behavior of liquidity premium in asset pricing,” *Journal of Financial Economics*, 34, 373-386.

Ellis, K., R. Michaely, and M. O’Hara, 1999, “The Accuracy of Trade Classification Rules: Evidence from the Nasdaq”, *Journal of Financial and Quantitative Analysis*, forthcoming.

Fama, E. F., and K. R. French, 1992, “The Cross-Section of Expected Stock Returns,” *Journal of Finance*, 47(2), 427-465.

Fama, E. F., and J. D. MacBeth, 1973, “Risk, Return, and Equilibrium: Empirical Tests,” *Journal of Political Economy*, 81, 607-636.

Jones, C. M., G. Kaul, M. L. Lipson, 1994, “Transactions, Volume, and Volatility,” *Review of Financial Studies*,” 7(4), 631-651.

Jones, C. M., and S. L. Slezak, 1999, “The Theoretical Implications of Asymmetric Information on the Dynamic and Cross-Sectional Characteristics of Asset Returns,” working paper, University of North Carolina - Chapel Hill.

Kyle, A. S., 1985, “Continuous Auctions and Insider Trading,” *Econometrica*, 53(6), 1315-1335.

Lee, C.M.C. and M.J. Ready, 1991, “Inferring Trade Direction from Intraday Data,” *Journal of Finance*, 46, 733-746.

Litzenberger, R., and C. Ramaswamy, 1979, “The Effect of Personal Taxes and Dividends on Capital Asset Prices: Theory and Empirical Evidence,” *Journal of Financial Economics*, 7, 163-196.

Loughran, T., 1997, “Book-to-Market across Firm Size, Exchange, and Seasonality: Is there an Effect?” *Journal of Financial and Quantitative Analysis*, 32(3), 249-268.

Miller, M. H., and M. S. Scholes , 1982, “Dividends and Taxes: Some Empirical Evidence” *Journal of Political Economy*, 90(6), 1118-1141.

O'Hara, M., 1995, *Market Microstructure Theory*. Blackwell Publishers.

Wang, J., 1993, “A Model of Intertemporal Asset Prices Under Asymmetric Information,” *Review of Economic Studies*, 60, 249-282.