Estimating the Order-Flow Component of Security Returns

Kerry Back, Kevin Crotty

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

Tao Li

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

Abstract

We derive a structural model of the amount of private information that is conveyed to the market via order flows. The model is a continuous-time Kyle model in which there is public information arrival and in which the strategic trader may or may not have private information. For empirical implementation, we treat each trading day as a separate instance of the model. Monte Carlo analysis shows that the key parameters are estimated well by maximum likelihood using intraday data over monthly time periods, even in the presence of some misspecification. We illustrate the procedure by estimating the model monthly for a single stock over a twenty-year time frame.

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

January 17, 2014

1. Introduction

What fraction of security returns is due to public information and what fraction is due to private information that is transmitted to the market via order flows? We develop and estimate a structural model of informed trading to answer this question. The model is a continuous-time Kyle (1985) model in which there may or may not be a private information event and in which there is public information arrival in addition to any information in order flows. The model can be estimated by maximum likelihood using discretely sampled price and order flow data. We use Monte Carlo analysis to assess the performance of the estimation procedure, and we implement the model with monthly estimation for a single stock over a twenty-year period.

The equilibrium pricing rule and trading strategy are obtained in closed form, up to evaluation of an integral. There is also a closed form expression for the market's conditional probability that an informed trader is present. The equilibrium aggregate order process is a generalization of a Brownian bridge. It appears as a Brownian motion to market makers. Their uncertainty about whether an informed trader is present, and about the signal of the informed trader should one be present, manifests itself as uncertainty about the ending point of the generalized Brownian bridge, causing it to be a Brownian motion relative to their information.

The model parameters of primary interest are the variance of private information (the probability of an information event multiplied by the variance of private information conditional on an event) and the fraction of the return variance that is due to private information. We treat each day as an independent realization of the theoretical model. We pool days to form monthly estimates of the parameters. We perform a Monte Carlo analysis to determine the reliability of the estimation procedure. The analysis shows that the parameters are estimated reasonably well over a time period

as short as one month when hourly price and order flow data are used but would be estimated much less reliably if only opening and closing prices and daily signed volume were used. We also use Monte Carlo analysis to assess how well the estimation procedure performs when the distribution of the private signal is misspecified. We find that it is robust to the misspecification we consider (a lognormal rather than triangular distribution).

As an example, we estimate the model using hourly price and order flow data over a twenty-year period for a single stock (Ashland, Inc.; ticker = ASH).[1] The mean estimate of the probability of an information event for ASH is 62%. The mean estimate of the fraction of the variance of daily ASH returns that is due to information in the order flow is 8.8%. The parameters appear roughly stationary over time, but there is a spike in the public information component of ASH returns during the financial crisis. The information in the order flow also increases during the crisis, but not to the same extent, causing the mean estimate of the fraction of returns due to order flow to drop to 3.4% during the crisis. This is consistent with information in the order flow being idiosyncratic and with systematic risk being relatively more important during the crisis.

The paper in the literature most closely related to ours is Odders-White and Ready (2008), who also estimate the probability and magnitude of an information event in the context of a Kyle (1985) model. In their model, as in ours, the signal of the informed trader is drawn from a mixture distribution (mixing over whether there was or was not an information event). However, Odders-White and Ready analyze a single-period model rather than a dynamic model. In a single-period model, because

[1]Our choice is a nod to the previous work of Easley, Kiefer and O'Hara (1997), who estimate their model using thirty trading days for ASH from 1990. ASH is also convenient because it traded continuously and underwent no stock splits during the sample period.

of the mixing, the conditional expectation of the asset value given the net order is not a linear function of the net order. To make the model tractable, Odders-White and Ready deviate from the usual Kyle model formulation and do not require the asset price to equal its conditional expected value. Instead, they only require that unconditional expected market maker profits are zero. They find the pricing rule that is linear in the net order that has this "zero conditional expected profits on average" property. Such a pricing rule would require commitment by market makers, because it is not consistent with ex-post optimization by market makers.

There are two advantages to analyzing a continuous-time Kyle model rather than a single period model. The first is tractability. Analytically solving a single period Kyle model in which there are zero conditional expected profits state by state seems to be impossible when the signal is drawn from a mixture distribution. However, as we demonstrate, it is feasible in continuous time. The second advantage is empirical. Like Odders-White and Ready, we estimate our model by assuming its duration corresponds to a day. Because we solve a dynamic Kyle model, we can use intraday data to obtain more efficient estimates. The Monte Carlo analysis shows large efficiency gains to using hourly rather than daily observations.

Our work is also related to the PIN model of Easley, Kiefer, O'Hara and Paperman (1996). A large literature in finance and accounting utilizes the PIN model to study information asymmetry in financial markets.[2] The PIN model estimates the probability of an information event using only order flow data.[3] Consequently, it does

[2]A portion of those papers assesses whether information risk is priced (e.g. Easley and O'Hara (2004), Duarte and Young (2009), Mohanram and Rajgopal (2009), Easley, Hvidkjaer and O'Hara (2002), Easley, Hvidkjaer and O'Hara (2010), Akins, Ng and Verdi (2012), Li, Wang, Wu and He (2009), Hwang, Lee, Lim and Park (2013)) while others use PIN as a measure of information asymmetry in a wide range of applications ranging from corporate finance (e.g. Chen, Goldstein and Jiang (2007), Ferreira and Laux (2007)) to accounting (e.g. Frankel and Li (2004), Jayaraman (2008)).

[3]In a paper introducing a concept called volume-synchronized PIN, Easley, Lopez de Prado and

not address the importance of information events. Our model is similar to the PIN model in that information events occur randomly, and we estimate the probability of an information event, treating each day as a separate instance of the model. However, we also use return data to assess the significance of information events.

Our work is also related to the literature that estimates the order-flow component of returns using linear regressions or vector auto-regressions of returns and order flows (see, for example, Glosten and Harris (1988), Hasbrouck (1991), Madhavan, Richardson and Roomans (1997), and Hasbrouck (2007)). These models presume that the coefficients that relate order flows to returns are time-invariant. In contrast, in our structural model, the relation between orders and prices is locally linear, but the linear coefficient (Kyle's lambda) is time-varying and depends on cumulative orders.

Our theoretical model generalizes Back (1992), principally by allowing for the possibility that there may be no private information. This causes the aggregate order process to be a generalization of a Brownian bridge (called a Doob h-transform – cf. Rogers and Williams (2000)) rather than a standard Brownian bridge as in Back (1992). A precursor to our paper is Li (2012), which also solves a continuous-time Kyle model in which the strategic trader may or may not be informed, by applying filtering theory to a transformation of the aggregate order process. The filtering solution produces a stochastic differential equation for the equilibrium rather than a closed form.

Related theoretical work includes Rossi and Tinn (2010), Banerjee and Green (2013), and Chakraborty and Yilmaz (2004). Rossi and Tinn solve a two-period Kyle model in which there are two large traders, one of whom is certainly informed and one

O'Hara (2012) use price changes over short time intervals to estimate the proportions of buys and sells. However, given the estimated order flows, they make no further use of prices. In contrast, we use both order flows and prices to estimate our model.

of whom may or may not be informed. In their model, unlike ours, there are always information events. Banerjee and Green solve a rational expectations model with myopic mean-variance investors, in which investors learn whether other investors are informed. They show that variation over time in the perceived likelihood of informed trading induces volatility clustering. While their model is quite different from ours, our model also exhibits volatility clustering. Volatility follows the same pattern as Kyle's lambda, which varies over time due to variation in the market's estimate of whether an information event occurred. Chakraborty and Yilmaz study a discrete-time Kyle model in which there may or may not be an information event. Their main result is that the informed trader will manipulate (sometimes buying when she has bad information and/or selling when she has good information) if the horizon is sufficiently long. The primary differences between their model and ours are that in their model the asset has only two possible values (high and low) and the noise trade distribution has finite support. If the low type trader never buys in their model, then an aggregate order larger than the maximum of the noise trade distribution implies for certain that the low type trader is not present. When the horizon is sufficiently long, it is optimal for the low type trader to deviate from a strategy of never buying and to buy until the aggregate order in a period is large enough that market makers put 100% probability on her not being a low type. Then, she can begin selling. Consequently, it cannot be an equilibrium in their model for low types to never buy and for high types to never sell, when the horizon is sufficiently long. In contrast, market makers in our model can never rule out any type of the informed trader until the end of the model, so it does not strictly pay for a low type to pretend to be a high type or vice versa (as we will show, and as is true in other continuous-time Kyle models, the informed trader in our model is locally indifferent about buying or selling, so pretending to be a different type is not suboptimal, but it does not occur

in equilibrium).

2. Theory

2.1. Model

Denote the time horizon for trading by $[0, 1]$. We identify it with a day in our empirical analysis. We assume the daily interest rate is small enough to be negligible. As is standard in Kyle models, we assume all traders are risk neutral. We assume that, in the absence of an information event, there are value traders present in the market who trade systematically against liquidity motivated traders, buying on price drops and selling on price rises. For simplicity, we assume the informed and value traders are organized as a monopoly and embodied in a single representative trader. The issue of competition among and between informed and value traders is interesting but beyond the scope of the paper.

Assume the single strategic trader receives a zero-mean signal S at time 0 with probability α. Let ξ denote an indicator for whether the strategic trader is informed ($\xi = 1$ if informed and $\xi = 0$ otherwise). Let V be a martingale that is independent of S and publicly observable. The value of the asset at the end of the day conditional on all available information is $V_1 + \xi S$. Assume the asset value becomes public information at the close of trading each day, and positions can then be liquidated frictionlessly. The standard continuous-time Kyle (1985) model is a special case of this model in which $\alpha = 1$, V is constant, and S is normally distributed.

In addition to the strategic trades, there are liquidity trades represented by a Brownian motion Z with zero drift and instantaneous standard deviation σ.[4] Let X_t denote the number of shares held by the strategic trader at date t (taking $X_0 = 0$

[4]In the proof of the proposition below, we assume that the Brownian motion Z is the projection map on $C[0, 1]$ equipped with Wiener measure. This ensures that the paths of Z are continuous with probability one even when we make a non-equivalent change of measure.

without loss of generality), and set $Y_t = X_t + Z_t$. The process Y is observed by market makers. Let $\mathcal{F}_t^{V,Y}$ denote the information of market makers at date t, which consists of the history of the processes V and Y.

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

$$P_t = \mathsf{E}\left[V_1 + \xi S \mid \mathcal{F}_t^{V,Y}\right] = V_t + \mathsf{E}\left[\xi S \mid \mathcal{F}_t^{V,Y}\right] . \tag{1}$$

We will show that there is an equilibrium in which $P_t = V_t + p(t, Y_t)$ for a function p. This means that the conditional expectation of ξS depends only on cumulative orders Y_t and not on the entire history of orders. The other requirement for equilibrium is that the strategic trades are optimal. Let θ_t denote the trading rate of the strategic trader (i.e., $\mathrm{d}X_t = \theta_t \, \mathrm{d}t$). The process θ has to be adapted to the information possessed by the strategic trader, which is V, ξS, and the history of Z (in equilibrium, the price reveals Z to the informed trader). The informed trader chooses the rate to maximize

$$\mathsf{E}\int_0^1 \left[V_1 + \xi S - P_t\right] \theta_t \, \mathrm{d}t = \mathsf{E}\int_0^1 \left[\xi S - p(t, Y_t)\right] \theta_t \, \mathrm{d}t \, , \tag{2}$$

with the function p being regarded by the informed trader as exogenous. In the optimization, we assume that the strategic trader is constrained to satisfy the "no doubling strategies" condition introduced in Back (1992), meaning that the strategy must be such that

$$\mathsf{E}\int_0^1 p(t, Y_t)^2 \, \mathrm{d}t < \infty$$

with probability one.

Assume the signal S has a continuous distribution function G. Set $\underline{s} = \inf\{s \mid G(s) > 0\}$ and $\bar{s} = \sup\{s \mid G(s) < 1\}$. Assume $-\infty \leq \underline{s} < 0 < \bar{s} \leq \infty$. Assume

G is strictly increasing on $(\underline{s}, \bar{s})$ except possibly on some interval containing zero. If there is such an interval with zero in its interior, then there is zero probability of very small good or bad news. Including this feature in the model would make it possible to ensure that information events are nontrivial. Under these assumptions, G^{-1} is uniquely defined on $(0, 1)$, except possibly at $G(0)$.

2.2. Brownian Bridge

Let F denote the distribution function of the normally distributed variable Z_1. Set $z_L = F^{-1}(\alpha G(0))$ and $z_H = F^{-1}(1 - \alpha + \alpha G(0))$. This means that

$$\alpha \operatorname{prob}(S \leq 0) = \operatorname{prob}(Z_1 \leq z_L)\,,$$

and

$$\alpha \operatorname{prob}(S > 0) = \operatorname{prob}(Z_1 > z_H)\,.$$

Thus, the unconditional probability of bad news is equal to the probability that $Z_1 \leq z_L$, and the unconditional probability of good news is equal to the probability that $Z_1 > z_H$.

Set

$$q(t, y, s) = \begin{cases} F^{-1}(\alpha G(s)) - y & \text{if } G(s) < G(0)\,, \\ \mathsf{E}[Z_1 \mid Z_t = y, z_L \leq Z_1 \leq z_H] - y & \text{if } G(s) = G(0)\,, \\ F^{-1}(1 - \alpha + \alpha G(s)) - y & \text{if } G(s) > G(0)\,. \end{cases} \tag{3}$$

Note that if $G(s) < G(0)$, then $z \stackrel{\text{def}}{=} F^{-1}(\alpha G(s))$ satisfies

$$F(z) = \alpha G(s) < \alpha G(0) = F(z_L)\,.$$

Thus, the function $s \mapsto F^{-1}(\alpha G(s))$ maps $\{s \mid G(s) < G(0)\}$ to $\{z \mid z < z_L\}$. Symmetrically, the function $s \mapsto F^{-1}(1 - \alpha + \alpha G(s))$ maps $\{s \mid G(s) > G(0)\}$ to

$\{z \mid z > z_H\}$.

Proposition. *Let* N *denote the standard normal distribution function. Let* $\mathbb{F}^Y = \{\mathcal{F}_t^Y \mid 0 \leq t \leq 1\}$ *denote the filtration generated by the stochastic process* Y *defined by* $Y_0 = 0$ *and*

$$\mathrm{d}Y_t = \frac{q(t, Y_t, \xi S)}{1-t}\,\mathrm{d}t + \mathrm{d}Z_t\,. \tag{4}$$

Then, the following are true:

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

(B) *With probability one,*

$$\xi = 1 \text{ and } S < 0 \quad \Rightarrow \quad Y_1 = F^{-1}(\alpha G(S)) < z_L\,, \tag{5a}$$

$$\xi = 0 \quad \Rightarrow \quad z_L \leq Y_1 \leq z_H\,, \tag{5b}$$

$$\xi = 1 \text{ and } S > 0 \quad \Rightarrow \quad Y_1 = F^{-1}(1 - \alpha + \alpha G(S)) > z_H\,. \tag{5c}$$

(C) *For each* $t < 1$*, the probability that* $\xi = 1$ *conditional on* $\mathcal{F}_t^Y$ *is*

$$\mathrm{N}\left(\frac{z_L - Y_t}{\sigma\sqrt{1-t}}\right) + 1 - \mathrm{N}\left(\frac{z_H - Y_t}{\sigma\sqrt{1-t}}\right)\,. \tag{6}$$

The process Y described in the proposition is a variation of a Brownian bridge. It differs from a Brownian bridge in that the endpoint is not uniquely determined when there is no information event ($\xi = 0$). Part (C) of the proposition follows immediately from the preceding parts, because the probability (6) is the probability that $Y_1 \notin [z_L, z_H]$ calculated on the basis that Y is an $\mathbb{F}^Y$–Brownian motion with zero drift and standard deviation σ.

2.3. Equilibrium

Let $f(\cdot \mid t, y)$ denote the density function of Z_1 conditional on $Z_t = y$, that is, the normal density function with mean y and variance $(1-t)\sigma^2$.

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

$$\theta_t = \frac{q(t, Y_t, \xi S)}{1-t}\,. \tag{7}$$

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

$$p(t,y) = \int_{-\infty}^{z_L} G^{-1}\left(\frac{F(z)}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z + \int_{z_H}^{\infty} G^{-1}\left(\frac{F(z)-1+\alpha}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z\,. \tag{8}$$

The asset price evolves as $\mathrm{d}P_t = \mathrm{d}V_t + \lambda(t, Y_t)\,\mathrm{d}Y_t$, where Kyle's lambda is

$$\begin{aligned}\lambda(t,y) = \frac{1}{\sigma^2(1-t)} &\int_{-\infty}^{z_L} (z-y) G^{-1}\left(\frac{F(z)}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z \\ &+ \frac{1}{\sigma^2(1-t)} \int_{z_H}^{\infty} (z-y) G^{-1}\left(\frac{F(z)-1+\alpha}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z\,. \end{aligned} \tag{9}$$

There is convergence to strong-form efficiency in the sense that $\lim_{t\to 1} P_t = V_1 + \xi S$ with probability one.

The probability that an information event occurred, conditional on the market's information at any date $t < 1$, is given by (6). The probability is generally an increasing function of the absolute net order imbalance at t; more precisely, it is an increasing function of the distance of the net order imbalance from the midpoint of z_L and z_H. The strong-form efficiency condition means that the market learns by the close of trading whether the strategic trader is informed and, if so, what her

information is. From the proposition, we know that if $\xi = 1$ and $S < 0$, then

$$Y_t \to F^{-1}(\alpha G(S)) < z_L \tag{10a}$$

with probability one as $t \to 1$. On the other hand, if $\xi = 1$ and $S > 0$, then

$$Y_t \to F^{-1}(1 - \alpha + \alpha G(S)) > z_H \tag{10b}$$

with probability one. In each case, the market learns S from Y as $t \to 1$. If the strategic trader is uninformed ($\xi = 0$), then

$$z_L \leq \liminf_{t \to 1} Y_t \leq \limsup_{t \to 1} Y_t \leq z_H \,, \tag{10c}$$

and the difference between P_t and V_t converges to zero as $t \to 1$.

The equilibrium is illustrated in Figure 1. The distribution of the signal S used to generate the figure is the same that we will use in our empirical work. Conditional on good news, the signal has a triangular distribution with lower limit 0, mode κ, and upper limit 2κ. The distribution for bad news is symmetric. See Section 3 for further discussion. Figure 1 shows that the conditional probability (6) of an information event is an increasing function of the absolute order imbalance Y_t. High trading in either direction is a signal to the market that an informed trader is present, as in PIN models. Kyle's lambda (9) is generally also an increasing function of the absolute order imbalance. The region in which it is decreasing (the outer tails) has low probability, being outside of a roughly two standard deviation band.[5] The standard deviation of price changes caused by order flows is proportional to Kyle's lambda,

[5]Two factors cause lambda to be decreasing in the absolute order imbalance in the tails. The first is that the conditional probability of an information event increases at a decreasing rate for high order imbalances; thus, further orders in the same direction have lower information content. The second is that the signal distribution is assumed to be bounded. If the market becomes quite sure there is an informed trader with an extreme signal, then adverse selection becomes low, because more extreme signals are unlikely or impossible.

because the standard deviation of order flows is assumed to be constant. Thus, there is stochastic volatility.

A sketch of the proof of the theorem is as follows. Market makers forecast the ending value of Y regarding it as a Brownian motion, and they estimate the value of S based on the Y forecast and based on the link (10) between the ending value of Y and S. This calculation produces the formula (8) for the equilibrium pricing function. In fact, the formula (8) is equivalent to

$$p(t, y) = \mathsf{E}[\pi(Y_1) \mid Y_t = y] \,, \tag{11}$$

where Y is taken to be a Brownian motion with zero drift and standard deviation σ as established in part (A) of the proposition, and where π is defined as

$$\pi(y) = \begin{cases} G^{-1}\left(\frac{F(y)}{\alpha}\right) & \text{if } y < z_L \,, \\ 0 & \text{if } z_L \leq y \leq z_H \,, \\ G^{-1}\left(\frac{F(y)-1+\alpha}{\alpha}\right) & \text{if } y > z_H \,. \end{cases} \tag{12}$$

Note that, by Part (B) of the proposition, $\pi(Y_1) = \xi S$ with probability one. Therefore, the price function (8) satisfies the equilibrium condition (1).

The process $p(t, Y_t)$ is a martingale given market makers' information due to the equilibrium condition (1). By Itô's formula and the fact that Y is a Brownian motion relative to market makers' information, the drift of $p(t, Y_t)$ relative to their information is

$$\frac{\partial}{\partial t} p(t, Z_t) + \frac{1}{2}\sigma^2 \frac{\partial^2}{\partial z^2} p(t, Z_t) \,.$$

Equating this to zero, Itô's formula yields

$$\mathrm{d}p(t, Y_t) = \frac{\partial p(t, Y_t)}{\partial y} \, \mathrm{d}Y_t \,.$$

Hence, Kyle's lambda equals $\partial p/\partial y$. The formula (9) for lambda follows from differentiating (8).

The strategic trader takes the pricing rule as given and optimizes over her trading rate. As in Back (1992), given the particular pricing rule (8), she is indifferent among all strategies that exhaust all profit opportunities in the sense of not leaving a gap between the ending price and the true value at the end of trading. Strategies that leave such a gap are clearly inferior. Therefore, the convergence results (10) established in the proposition imply the optimality of the strategy (7). A verification theorem is provided in the appendix.

3. Maximum Likelihood Estimation

Assume the trading period $[0, 1]$ corresponds to a day. This implies that any private information becomes public before trading opens on the following day.[6] We can estimate the model parameters using intraday price and order flow information. If we assume further that the model parameters are stable over time, then the price and order flow information from multiple days can be merged to estimate the parameters with greater precision.[7]

The opening price on each day i is $P_{i0} \stackrel{\text{def}}{=} \mathsf{E}[V_{i1} + \xi_i S_i] = V_{i0}$. To obtain stationarity, we assume that the signal S_i on day i is proportional to the observed opening price P_{i0}. To be precise, we assume that S_i/P_{i0} is a sequence of iid random variables with common distribution function G. This construction causes the pricing

[6] In contrast to Odders-White and Ready (2008), our estimation does not use overnight returns. In our theoretical model, private information that is made public after the close of trading is incorporated into prices before trading ends (convergence to strong-form efficiency). Thus, overnight returns in our model are due to arrival of new public information, which does not aid in estimating the model.

[7] We estimate the model at a monthly frequency, so concerns about parameter instability should be lower than with the annual estimation window used in, for example, Easley, Hvidkjaer and O'Hara (2002), Odders-White and Ready (2008), and Duarte and Young (2009).

function to be day-specific, and we denote it by $p_i(t, y)$. In fact,

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

where $p(t, y)$ is defined in the theorem in terms of the distribution function G. We specify the function G used in our empirical implementation below.

The price at time t on day i is $V_{it} + p_i(t, Y_{it})$, so the gross return through time t is

$$\frac{P_{it}}{P_{i0}} = \frac{V_{it}}{V_{i0}} + \frac{p_i(t, Y_{it})}{P_{i0}} = \frac{V_{it}}{V_{i0}} + p(t, Y_{it})\,. \tag{13}$$

Assume

$$\frac{\mathrm{d}V_{it}}{V_{it}} = \delta\,\mathrm{d}B_{it}$$

for a constant δ and a Brownian motion B_i, so we have

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

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

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

On each day i, the vector $Y_i = (Y_{i,t_1}, \ldots, Y_{i,t_{k+1}})'$ is normally distributed with

mean 0 and covariance matrix $\sigma^2\Delta\Sigma$. Set

$$U_{ij} = \log\left(\frac{P_{ij}}{P_{i0}} - p(t_j, Y_{ij})\right) \tag{14}$$

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

$$f(U_{i1}, \ldots U_{i,k+1})\mathrm{e}^{-\sum_{j=1}^{k+1} U_{ij}}\,,$$

where f denotes the multivariate normal density function with mean vector $-(\delta^2\Delta/2)\Gamma$ and covariance matrix $\delta^2\Delta\Sigma$.

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

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

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

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

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

$$\begin{aligned} -\mathcal{L} = n(k+1)\log\sigma + \frac{1}{2\sigma^2\Delta}\sum_{i=1}^{n} Y_i'\Sigma^{-1}Y_i + n(k+1)\log\delta \\ + \frac{1}{2\delta^2\Delta}\sum_{i=1}^{n} U_i'\Sigma^{-1}U_i + \frac{n\delta^2}{8} + \sum_{i=1}^{n}\left(\sum_{i=1}^{k} U_{ik} + \frac{3}{2}U_{i,k+1}\right)\,. \end{aligned} \tag{15}$$

This can be minimized analytically in δ. The partial derivative of $-\mathcal{L}$ with respect

to δ is

$$\frac{n(k+1)}{\delta} + \frac{n\delta}{4} - \frac{1}{\delta^3 \Delta} \sum_{i=1}^{n} U_i' \Sigma^{-1} U_i \,, \tag{16}$$

so the first-order condition is

$$n(k+1)\delta^2 + \frac{n\delta^4}{4} - \frac{1}{\Delta} \sum_{i=1}^{n} U_i' \Sigma^{-1} U_i = 0 \,. \tag{17}$$

This is a quadratic equation in δ^2 with unique positive solution

$$\delta^2 = -2(k+1) + 2\sqrt{(k+1)^2 + \frac{1}{n\Delta} \sum_{i=1}^{n} U_i' \Sigma^{-1} U_i} \,. \tag{18}$$

The sign of the partial derivative (16) for $\delta > 0$ is the sign of the left-hand side of (17), which is increasing in δ and therefore positive to the right of (18) and negative to the left. Hence, (18) is the unique minimizer of $-\mathcal{L}$. Substituting (18) into (15) gives the reduced log-likelihood function, which we maximize in α and σ and in the parameters of the distribution function G (which enter $\mathcal{L}$ because they affect the function p that enters $\mathcal{L}$ via (14)).

The signal is assumed in the model to have a zero mean (without loss of generality, because the mean can be embedded in V_0). Therefore, in the interest of parsimony, we use a one-parameter distribution for the signal in the estimation, with the parameter being a scale parameter. The theoretical model can accommodate an interval of zero probability around a signal value of zero. We do not exploit that generality, but we do assume that the density of the signal distribution is zero at zero. Specifically, let

$$S_i = P_{i0} \chi_i W_i \,,$$

where $\chi_i \in \{-1, 1\}$ with equal probabilities and where W_i is drawn from a triangular

distribution with lower limit 0, mode κ, and upper limit 2κ. Then,

$$G(s) = \begin{cases} 0 & \text{if } s \leq -2\kappa\,, \\ \frac{(2\kappa+s)^2}{4\kappa^2} & \text{if } -2\kappa < s \leq -\kappa\,, \\ \frac{1}{2} - \frac{s^2}{4\kappa^2} & \text{if } -\kappa < s < 0\,, \\ \frac{1}{2} & \text{if } s = 0\,, \\ \frac{1}{2} + \frac{s^2}{4\kappa^2} & \text{if } 0 < s \leq \kappa\,, \\ 1 - \frac{(2\kappa-s)^2}{4\kappa^2} & \text{if } \kappa < s \leq 2\kappa\,, \\ 1 & \text{if } 2\kappa < s\,. \end{cases}$$

We show in the next section that the estimate of the fraction of return variance due to order flow information is robust to misspecification of this signal distribution.

4. Monte Carlo

To assess the efficacy of the maximum likelihood procedure and to determine the importance of using intraday data, we perform a Monte Carlo analysis. We simulate hourly and closing prices and net order imbalances ($\Delta = 1/6.5$ and $k = 6$) for 22 days and estimate the model parameters. We repeat this 1000 times.[8] There are two combinations of the parameters in which we are particularly interested. The first is the standard deviation of the signal ξS. This equals

$$\sqrt{\frac{7\alpha}{6}}\kappa\,. \tag{19}$$

[8]We use the Nelder-Mead algorithm to maximize the log-likelihood function. Each month, we initialize both κ and δ as the sample standard deviation of that month's log open-to-close price changes. The volatility of order imbalances, σ, is initialized at the sample standard deviation of the month's closing order imbalances. We run the algorithm with various initial values for α ($\alpha \in \{0.25, 0.5, 0.75\}$) and choose the optimum.

The second combination is the proportion of the return variance that is due to private information. This is equal to

$$\frac{\text{var}(\xi S)}{\text{var}(\xi S) + \text{var}(\text{e}^{\delta B_{i1} - \delta^2/2})} = \frac{\alpha\kappa^2}{\alpha\kappa^2 + 6\left(\text{e}^{\delta^2} - 1\right)/7}\,. \tag{20}$$

The Monte Carlo results are reported in Figures 2 and 3. The true parameter values are displayed in the figures (indicated by the dashed vertical lines). Clearly, there is substantial benefit to using intraday data. The standard deviations of the estimators are reduced substantially by using intraday data. The mean estimates of the probability of an information event α, the signal scale parameter κ, and the fraction (20) of the return variance due to order flow information are substantially higher than the true values when only daily observations are used but are much closer to the true values when intraday data is used. These improvements are reflected by a reduction of approximately 70% in root mean square errors for the signal standard deviation and the order flow component of returns.

We estimate the variance-covariance matrix of the estimates using the outer-product method (Hamilton, 1994).[9] The standard errors are quite accurate for the parameters (19) and (20) that are of primary interest: the 90% confidence interval contains the true parameter value in 90.3% and 89.4% of the simulations, respectively.[10] The standard errors are also quite accurate for σ and δ, for which the true values lie within the confidence intervals in 92.0% and 91.5% of the simulations, respectively. The standard errors are less accurate for α and κ, the true values of which fall within the 90% confidence interval in 61.7% and 85.3% of the simulations, respectively.

[9] Under this method, the estimate of the information matrix is the average daily outer-product of the gradient of $\mathcal{L}_i$, evaluated at the estimated parameter vector. The derivatives are calculated numerically.

[10] We estimate standard errors for (19) and (20) using the delta method.

We also test the robustness of our estimation procedure to a different underlying signal distribution. To this end, we simulate hourly and closing prices and net order imbalances ($\Delta = 1/6.5$ and $k = 6$) for 22 days using a lognormal signal distribution with $W_i = \kappa_1 \exp(\kappa_2 \varepsilon_i - 0.5\kappa_2^2)$ with $\kappa_1 = 0.5\%$, $\kappa_2 = 0.25\%$, and standard normal ε_i. Again, we repeat 1000 times. The Monte Carlo results are reported in Figure 4. While α is generally overestimated, the estimation procedure does reasonably well capturing the parameters (19) and (20), with root mean square errors quite close to those obtained with a correctly specified signal distribution. Misspecification of the signal distribution also does not lead to large size distortions: 90% confidence intervals contain the true signal standard deviation and order flow component values in 91.9% and 90.5% of the simulations, respectively.

5. Empirics

We demonstrate the estimation procedure for a single firm, Ashland Incorporated (ticker: ASH). We use TAQ data on stock trade and quote information for ASH from 1993 through 2012. We sign trades as buys and sells using the Lee and Ready (1991) algorithm: trades above (below) the prevailing quote midpoint are considered buys (sells). If a trade occurs at the midpoint, then the trade is classified as a buy (sell) if the trade price is greater (less) than the previous differing transaction price. Order imbalance is defined as shares bought less shares sold, expressed in units of one million shares.

Open-to-close summary statistics for daily net returns, order flows, and volume are shown in Table 1 for odd years in the sample. Total trading volume has increased sharply for ASH, a trend representative of the broader market. However, order imbalances appear stationary over the sample.

We sample the trading day on the half-hour and at the close, so $k = 6$ with

$\Delta = 1/6.5$. We estimate the model monthly using the same numerical method that we use for the Monte Carlo analysis (see footnote 8). Figure 5 plots the histograms of the monthly estimates. Key aspects of the estimates are that the average estimate of the probability of an information event is 62%, and the average estimate of the order flow component of returns is 9%.

Figure 6 plots the time series of monthly estimates. The parameters appear to be roughly stationary. There were spikes in both private information κ and public information δ during the financial crisis. The increase in public information was especially pronounced, which led to a relatively low order flow component of returns during the crisis. The mean estimate of the fraction of returns due to order flow drops to 3.4% during the crisis. As noted before, this is consistent with information in the order flow being idiosyncratic and with systematic risk being relatively more important during the crisis. The volatility of liquidity trading experienced a spike in 2000, around the time of the dot-com crash. Despite the well-documented rise of high-frequency trading and the associated sharp increase in trading volume, the estimated volatility of order imbalances has remained quite stable over the twenty year sample, consistent with the summary statistics above.

We estimate the variance-covariance matrix of the estimates using the outer-product method as in the Monte Carlo analysis.[11] The 90% confidence intervals are plotted in Figure 6. The estimates of σ and δ are quite precise while the estimates of α and κ are noisier. In particular, the latter are less precisely estimated in the second half of the sample. We estimate standard errors for (19) and (20) using the delta method. For the first half of the sample, both the signal standard deviation and the

[11] See footnote 9. If the estimates of κ or α are very close to zero, the likelihood is invariant to the value of the other parameter. This is the case for ten firm-months due to low κ estimates. In these instances, we set the standard error of the α estimate equal to one.

order flow component of returns are statistically different from zero in general. The signal standard deviation is also statistically significant throughout the recent crisis.

6. Conclusion

We derive and estimate a structural model of the amount of private information that is conveyed to the market via order flows. The model is a dynamic Kyle model in which the strategic trader may or may not have private information and in which there is also public information arrival. We show how to estimate the model by maximum likelihood. Monte Carlo analysis demonstrates large efficiency gains to using intraday data. Monte Carlo analysis also shows that the estimate of the proportion of returns due to order flow information is robust to misspecification of the distribution of private information. We illustrate the method by estimating the parameters monthly using intraday price and order flow data for a single stock over a twenty-year time frame. On average, approximately 9% of the daily return variance is due to information in order flows.

Appendix A. Proofs

Take the set of states of the world to be $\Omega \stackrel{\text{def}}{=} \mathbb{R} \times C[0,1]$, where $C[0,1]$ denotes the set of continuous functions on $[0,1]$. The probability distribution $\mathbb{P}$ on Ω (denoted as prob in the text) is the product of the distribution of $\xi S \in \mathbb{R}$ with Wiener measure on $C[0,1]$. Let $\mathbb{G} \equiv \{\mathcal{G}_t \mid 0 \leq t \leq 1\}$ denote the product of the trivial σ–field on $\mathbb{R}$ (namely, $\{\emptyset, \mathbb{R}\}$) with the usual complete filtration on $C[0,1]$, and let $\mathbb{F} = \{\mathcal{F}_t \mid 0 \leq t \leq 1\}$ denote the product of the Borel σ–field on $\mathbb{R}$ with the same filtration on $C[0,1]$. The filtration $\mathbb{F}$ represents the informed trader's information. The Brownian motion Z is the usual projection map on $C[0,1]$; more precisely, $Z_t(\omega) = z_t$, where $\omega = (a, z) \in \mathbb{R} \times C[0,1]$. The completed filtration generated by Z is $\mathbb{G}$.

Proof of the Proposition. As stated earlier, $f(\cdot \mid t, y)$ denotes the normal density function with mean y and standard deviation $\sigma\sqrt{1-t}$. Set

$$k(t,y,s) = \begin{cases} f(F^{-1}(\alpha G(s)) \mid t, y) & \text{if } s < 0\,, \\ \int_{z_L}^{z_H} f(z \mid t, y)\,\mathrm{d}z & \text{if } s = 0\,, \\ f(F^{-1}(1-\alpha+\alpha G(s)) \mid t, y) & \text{if } s > 0\,. \end{cases}$$

Define

$$\ell(t,y,s) = \frac{\partial \log k(t,y,s)}{\partial y}\,.$$

Then,

$$\begin{aligned}
s < 0 \quad &\Rightarrow \quad (1-t)\sigma^2 \ell(t,y,s) = F^{-1}(\alpha G(s)) - y\,, \\
s = 0 \quad &\Rightarrow \quad (1-t)\sigma^2 \ell(t,y,s) = \frac{\int_{z_L}^{z_H} (z-y) f(z \mid t,y)}{\int_{z_L}^{z_H} f(z \mid t,y)} \\
&\qquad\qquad\qquad\qquad = \mathsf{E}\left[Z_1 - y \mid Z_t = y, z_L \le Z_1 \le z_H\right]\,, \\
s > 0 \quad &\Rightarrow \quad (1-t)\sigma^2 \ell(t,y,s) = F^{-1}(1-\alpha+\alpha G(s)) - y\,.
\end{aligned}$$

Thus, $(1-t)\sigma^2 \ell(t,y,s) = q(t,y,s)$, and the stochastic differential equation (4) can be written as

$$\mathrm{d}Y_t = \sigma^2\, \ell(t, Y_t, \xi S)\, \mathrm{d}t + \mathrm{d}Z_t \tag{A.1}$$

To put this in a more standard form, define the two-dimensional process $\hat{Y}_t = (\xi S, Y_t)$ with random initial condition $\hat{Y}_0 = (\xi S, 0)$, and augment (A.1) with the equation $\mathrm{d}\xi S = 0$. The existence of a unique strong solution $\hat{Y}$ to this enlarged system follows from Lipschitz and growth conditions satisfied by ℓ. See Karatzas and Shreve (1988, Theorem 5.2.9).

The uniqueness in distribution of weak solutions of stochastic differential equations (Karatzas and Shreve, 1988, Theorem 5.3.10) implies that we can demonstrate Properties (A) and (B) by exhibiting a weak solution for which they hold. To construct such a weak solution, define a new measure $\mathbb{Q}_\tau$ on each σ–field $\mathcal{F}_\tau$ for each deterministic time $\tau < 1$ using $k(\tau, Z_\tau, \xi S)/k(0,0,\xi S)$ as the Radon-Nikodym derivative. Because $f(z \mid \cdot)$ satisfies the Kolmogorov backward equation

$$\frac{\partial f(z \mid t,y)}{\partial t} + \frac{\sigma^2}{2} \frac{\partial^2 f(z \mid t,y)}{\partial y^2} = 0\,,$$

it follows from Itô's formula and the regularity of k that $k(t, Z_t, \xi S)$ restricted to $[0, \tau]$

is a martingale on the filtration $\mathbb{F}$. Consequently, the measures Q_τ are probability measures and are consistent in the sense that $\mathbb{Q}_\tau$ restricted to $\mathcal{F}_{\tau'}$ equals $\mathbb{Q}_{\tau'}$ for $\tau' < \tau$. The filtration $\mathbb{F}$ is left-continuous at $t = 1$, meaning that $\mathcal{F}_1$ is the smallest σ–field containing $\cup_{t<1}\mathcal{F}_t$. Caratheodory's Extension Theorem therefore implies the existence of a unique measure $\mathbb{Q}$ on $\mathcal{F}_1$ such that $\mathbb{Q}(A) = \mathbb{Q}_\tau(A)$ if $A \in \mathcal{F}_\tau$ for any $\tau < 1$. By Girsanov's Theorem, the process Z^* defined by $Z_0^* = 0$ and

$$\mathrm{d}Z_t^* = -\sigma^2\,\ell(t, Z_t, \xi S)\,\mathrm{d}t + \mathrm{d}Z_t$$

is a Brownian motion (with zero drift and standard deviation σ) on the filtration $\mathbb{F}$ relative to $\mathbb{Q}$ (Revuz and Yor, 1991, Theorem VIII.1.4). It follows that Z is a weak solution of (A.1) relative to the Brownian motion Z^* on the filtered probability space $(\Omega, \mathbb{F}, \mathbb{Q})$.

To establish Property (A) for the weak solution, we need to show that Z is a Brownian motion on $(\Omega, \mathbb{G}, \mathbb{Q})$. Because Z is a Brownian motion on $(\Omega, \mathbb{G}, \mathbb{P})$, it suffices to show that $\mathbb{Q} = \mathbb{P}$ when both are restricted to $\mathcal{G}_1$. Because $\mathbb{G}$ is left-continuous, it actually suffices to show that $\mathbb{Q} = \mathbb{P}$ when both are restricted to $\mathcal{G}_\tau$ for arbitrary $\tau < 1$. This holds if for all $t_1 < \cdots < t_n \leq \tau$ and all Borel B we have

$$\mathbb{P}((Z_{t_1}, \ldots, Z_{t_n}) \in B) = \mathbb{Q}((Z_{t_1}, \ldots, Z_{t_n}) \in B)\,. \tag{A.2}$$

The right-hand side of (A.2) equals

$$\mathsf{E}\left[\frac{k(t_n, Z_{t_n}, \xi S)}{k(0, 0, \xi S)}\mathbf{1}_B(Z_{t_1}, \ldots, Z_{t_n})\right], \tag{A.3}$$

which can be represented as the following sum:

$$\mathsf{E}\left[1_{\{\xi S<0\}}\frac{k(t_n,Z_{t_n},S)}{k(0,0,S)}\mathbf{1}_B(Z_{t_1},\ldots,Z_{t_n})\right]+\mathsf{E}\left[1_{\{\xi S=0\}}\frac{k(t_n,Z_{t_n},0)}{k(0,0,0)}\mathbf{1}_B(Z_{t_1},\ldots,Z_{t_n})\right]$$
$$+\mathsf{E}\left[1_{\{\xi S>0\}}\frac{k(t_n,Z_{t_n},S)}{k(0,0,S)}\mathbf{1}_B(Z_{t_1},\ldots,Z_{t_n})\right] \quad \text{(A.3}')$$

To evaluate the first term in (A.3′), define $\tilde{z}=F^{-1}(\alpha G(\xi S))$. Then,

$$1_{\{\xi S<0\}}\frac{k(t_n,Z_{t_n},S)}{k(0,0,S)}=1_{\{\tilde{z}<z_L\}}\frac{f(\tilde{z}\mid t_n,Z_{t_n})}{f(\tilde{z}\mid 0,0)}\,. \tag{A.4}$$

Furthermore, for any $z<z_L$, the formula $\tilde{z}=F^{-1}(\alpha G(\xi S))$ implies that

$$\mathbb{P}(\tilde{z}\le z)=\alpha\,\mathbb{P}\left(S\le G^{-1}\left(\frac{F(z)}{\alpha}\right)\right)\,.$$

It follows that the $\mathbb{P}$–density function of $\tilde{z}$ on the interval $(-\infty,z_L)$ is the derivative of F (that is, the $\mathbb{P}$–density function of Z_1) which is $f(\cdot\mid 0,0)$ in our notation. Hence, the first term in (A.3′) is

$$\int_B \mathrm{d}z_{t_1}\cdots\mathrm{d}z_{t_n}\,f_{t_1\cdots t_n}(z_{t_1},\ldots,z_{t_n})\int_{-\infty}^{z_L}\mathrm{d}z\,f(z\mid t_n,z_{t_n})\,,$$

where $f_{t_1\cdots t_n}$ denotes the joint $\mathbb{P}$–density function of $(Z_{t_1},\ldots,Z_{t_n})$. This equals

$$\mathbb{P}((Z_{t_1},\ldots,Z_{t_n})\in B,Z_1<z_L)\,. \tag{A.5a}$$

An analogous calculation based on the definition $\tilde{z}=F^{-1}(1-\alpha+\alpha G(\xi S))$ shows that the third term in (A.3′) equals

$$\mathbb{P}((Z_{t_1},\ldots,Z_{t_n})\in B,Z_1>z_H)\,. \tag{A.5b}$$

The middle term in (A.3′) equals

$$\mathsf{E}\left[1_{\{\xi S=0\}}\frac{\int_{z_L}^{z_H} f(z \mid t_n, Z_{t_n})\,\mathrm{d}z}{\int_{z_L}^{z_H} f(z \mid 0,0)\,\mathrm{d}z}1_B(Z_{t_1},\ldots,Z_{t_n})\right].$$

Using the $\mathbb{P}$–independence of ξS and Z and the fact that

$$\mathsf{E}\left[1_{\{\xi S=0\}}\right] = \int_{z_L}^{z_H} f(z \mid 0,0)\,\mathrm{d}z = 1-\alpha\,,$$

this simplifies to

$$\int_B \mathrm{d}z_{t_1}\cdots\mathrm{d}z_{t_n}\, f_{t_1\cdots t_n}(z_{t_1},\ldots,z_{t_n})\int_{z_L}^{z_H}\mathrm{d}z\, f(z \mid t_n, z_{t_n})\,,$$

which equals

$$\mathbb{P}((Z_{t_1},\ldots,Z_{t_n})\in B, z_L\le Z_1\le z_H)\,. \tag{A.5c}$$

Adding (A.5a)–(A.5c) yields (A.2).

To establish Property (B) for the weak solution of (A.1), we need to show that

$$\mathbb{Q}\left(Z_1 = F^{-1}(\alpha G(\xi S)) \mid \xi S<0\right)=1\,, \tag{A.6a}$$

$$\mathbb{Q}\left(Z_1\in[z_L,z_H] \mid \xi S=0\right)=1\,, \tag{A.6b}$$

$$\mathbb{Q}\left(Z_1 = F^{-1}(1-\alpha+\alpha G(\xi S)) \mid \xi S>0\right)=1\,. \tag{A.6c}$$

We will use the fact that, for any $\gamma>0$,

$$f(z \mid t,y)1_{\{|z-y|>\gamma\}} \le \frac{1}{\sqrt{2(1-t)\sigma^2\pi}}\mathrm{e}^{-\frac{\gamma^2}{2(1-t)\sigma^2}}\to 0 \tag{A.7}$$

as $t\to 1$.

Consider (A.6a). Set $\tilde{z}=F^{-1}(\alpha G(\xi S))$. Select an arbitrary $\gamma>0$. Consider the

event $|Z_t - \tilde{z}| > \gamma$ for any $t < 1$. From (A.4) and the fact that the $\mathbb{P}$–density function of $\tilde{z}$ is $f(\cdot \mid 0, 0)$, we obtain

$$\begin{aligned}\mathbb{Q}\left(|Z_t - \tilde{z}| > \gamma \mid \xi S < P_0\right) &= \mathsf{E}\left[\frac{f(\tilde{z} \mid t, Z_t)}{f(\tilde{z} \mid 0, 0)} 1_{\{|Z_t - \tilde{z}| > \gamma\}} \,\middle|\, \tilde{z} < z_L\right] \\ &= \frac{2}{\alpha} \int_{-\infty}^{\infty} \mathrm{d}z_t \, f_t(z_t) \int_{-\infty}^{z_L} \mathrm{d}z \, f(z \mid t, z_t) 1_{\{|z_t - z| > \gamma\}} \,,\end{aligned}$$

where f_t denotes the $\mathbb{P}$–density function of Z_t. This converges to 0 as $t \to 1$, by (A.7). Thus, Z_t converges in $\mathbb{Q}$–probability to $\tilde{z}$ as $t \to 1$, conditional on $\xi S < P_0$. This implies that a subsequence converges with $\mathbb{Q}$–probability one. However, Z has continuous paths, so $Z_1 = \tilde{z}$ with $\mathbb{Q}$–probability one when $\xi S < 0$. An analogous argument establishes (A.6c).

To verify (A.6b), note that

$$\mathbb{Q}\left(Z_t < z_L - \gamma, \xi S = 0\right) = \mathsf{E}\left[1_{\{\xi S = 0\}} \frac{\int_{z_L}^{z_H} f(z \mid t, Z_t) \, \mathrm{d}z}{\int_{z_L}^{z_H} f(z \mid 0, 0) \, \mathrm{d}z} 1_{\{Z_t < z_L - \gamma\}}\right] .$$

Using the $\mathbb{P}$–independence of ξS and Z and the fact that

$$\mathsf{E}\left[1_{\{\xi S = 0\}}\right] = \int_{z_L}^{z_H} f(z \mid 0, 0) \, \mathrm{d}z = 1 - \alpha \,,$$

this simplifies to

$$\int_{-\infty}^{\infty} \mathrm{d}z_t \, f_t(z_t) \int_{z_L}^{z_H} \mathrm{d}z \, f(z \mid t, z_t) 1_{\{z_t < z_L - \gamma\}} \,,$$

which converges to 0 as $t \to 1$ by virtue of (A.7). Likewise, $\mathbb{Q}\left(Z_t > z_H + \gamma, \xi S = P_0\right) \to 0$ as $t \to 1$. Due to the path continuity of Z, this implies (A.6b).

□

Proof of the Theorem. It is explained in the text why the equilibrium condition (1) holds and why the formula given for Kyle's lambda is correct. It remains to show that the strategy (7) is optimal for the informed trader. The value function J of the informed trader depends on time, on ξS, and on Y_t. We begin by defining it at time $t = 1$. For each $s < 0$, set $\omega(s) = F^{-1}(\alpha G(s))$. For $s > 0$, define $\omega(s) = F^{-1}(1-\alpha+\alpha G(s))$. For $s = 0$, let $\omega(s)$ denote a fixed but arbitrary point in $[z_L, z_H]$. Set

$$J(1, y, s) = \int_y^{\omega(s)} (s - \pi(z))\,\mathrm{d}z\,. \tag{A.8}$$

This definition is invariant to the choice of $\omega(0) \in [z_L, z_H]$, because $\pi(z) = 0$ for $z \in [z_L, z_H]$.

We have $J(1, y, s) \geq 0$ with equality if and only if $\pi(y) = s$. Furthermore

$$\frac{\partial J(1, y, s)}{\partial y} = \pi(y) - s\,.$$

For $t < 1$, set

$$J(t, y, s) = \mathsf{E}[J(1, Z_1, s) \mid Z_t = y] = \int_{-\infty}^{\infty} J(1, z, s) f(z \mid t, y)\,\mathrm{d}z\,. \tag{A.9}$$

This construction implies that $J(t, Z_t, s)$ is a martingale for each fixed s, so it has zero drift. From Itô's formula, its drift is

$$\frac{\partial}{\partial t} J(t, Z_t, s) + \frac{\sigma^2}{2} \frac{\partial^2}{\partial z^2} J(t, Z_t, s)\,.$$

Equating this to zero, Itô's formula implies

$$J(1, Y_1, \xi S) = J(0, 0, \xi S) + \int_0^1 \mathrm{d}J(t, Y_t, \xi S) = J(0, 0, \xi S) + \int_0^1 \frac{\partial J(t, Y_t, \xi S)}{\partial y}\,\mathrm{d}Y_t\,.$$

Therefore,

$$\mathsf{E}[J(1, Y_1, \xi S) - J(0, 0, \xi S)] = \mathsf{E} \int_0^1 \frac{\partial J(t, Y_t, \xi S)}{\partial y} \, \mathrm{d}Y_t \,. \tag{A.10}$$

To calculate $\partial J(t, y, s)/\partial y$, rewrite (A.9) as

$$J(t, y, s) = \int_{-\infty}^{\infty} J(1, y + z, s) f(z \mid t, 0) \, \mathrm{d}z \,.$$

From this and (A.8), we obtain

$$\begin{aligned} \frac{\partial J(t, y, s)}{\partial y} &= \int_{-\infty}^{\infty} \frac{\partial J(1, y + z, s)}{\partial y} f(z \mid t, 0) \, \mathrm{d}z \\ &= \int_{-\infty}^{\infty} [\pi(y + z) - s] f(z \mid t, 0) \, \mathrm{d}z \\ &= p(t, y) - s \,, \end{aligned}$$

using (11) for the last equality. Therefore,

$$\frac{\partial J(t, Y_t, \xi S)}{\partial y} = p(t, Y_t) - \xi S \,.$$

Combining this with (A.10) and using the fact that $Y = X + Z$ yields

$$\mathsf{E}[J(1, Y_1, \xi S) - J(0, 0, \xi S)] = \mathsf{E} \int_0^1 [p(t, Y_t) - \xi S] \theta_t \, \mathrm{d}t + \mathsf{E} \int_0^1 [p(t, Y_t) - \xi S] \, \mathrm{d}Z_t \,.$$

The "no doubling strategies" condition implies that $\int p \, \mathrm{d}Z$ is a martingale, so the second term on the right-hand side is zero. Rearranging gives

$$\mathsf{E} \int_0^1 [\xi S - p(t, Y_t)] \theta_t \, \mathrm{d}t = \mathsf{E}[J(0, 0, \xi S) - J(1, Y_1, \xi S)] \leq \mathsf{E}[J(0, 0, \xi S)] \,.$$

Thus, $\mathsf{E}[J(0, 0, \xi S)]$ is an upper bound on the expected profit, and the bound is

achieved if and only if $\mathsf{E}[J(1, Y_1, \xi S)] = 0$, which is equivalent to $\pi(Y_1) = \xi S$ with probability one. By part (B) of the proposition, the strategy (7) is therefore optimal.

□

References

Akins, B., Ng, J., Verdi, R.S., 2012. Investor competition over information and the pricing of information asymmetry. The Accounting Review 87, 35–58.

Back, K., 1992. Insider trading in continuous time. Review of Financial Studies 5, 387–409.

Banerjee, S., Green, B., 2013. Learning whether other traders are informed. Working Paper.

Chakraborty, A., Yilmaz, B., 2004. Manipulation in market order models. Journal of Financial Markets 7, 187–206.

Chen, Q., Goldstein, I., Jiang, W., 2007. Price informativeness and investment sensitivity to stock price. Review of Financial Studies 20, 619–650.

Duarte, J., Young, L., 2009. Why is PIN priced? Journal of Financial Economics 91, 119–138.

Easley, D., Hvidkjaer, S., O'Hara, M., 2002. Is information risk a determinant of asset returns. Journal of Finance 57, 2185–2221.

Easley, D., Hvidkjaer, S., O'Hara, M., 2010. Factoring information into returns. Journal of Financial and Quantitative Analysis 45, 293–309.

Easley, D., Kiefer, N.M., O'Hara, M., 1997. One day in the life of a very common stock. Review of Financial Studies 10, 805–835.

Easley, D., Kiefer, N.M., O'Hara, M., Paperman, J.B., 1996. Liquidity, information, and infrequently traded stocks. Journal of Finance 51, 1405–1436.

Easley, D., O'Hara, M., 2004. Information and the cost of capital. Journal of Finance 59, 1553–1583.

Easley, D., Lopez de Prado, M., O'Hara, M., 2012. Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies 25, 1457–1493.

Ferreira, M.A., Laux, P.A., 2007. Corporate governance, idiosyncratic risk, and information flow. Journal of Finance 62, 951–989.

Frankel, R., Li, X., 2004. Characteristics of a firm's information environment and the information asymmetry between insiders and outsiders. Journal of Accounting and Economics 37, 229–259.

Glosten, L.R., Harris, L.E., 1988. Estimating the components of the bid/ask spread. Journal of Fianncial Econmics 21, 123–142.

Hamilton, J., 1994. Time Series Analysis. Princeton University Press.

Hasbrouck, J., 1991. Measuring the information content of stock trades. Journal of Finance 46, 179–207.

Hasbrouck, J., 2007. Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Security Trading. Oxford University Press, Oxford.

Hwang, L.S., Lee, W.J., Lim, S.Y., Park, K.H., 2013. Does information risk affect the implied cost of equity capital? an analysis of PIN and adjusted PIN. Journal of Accounting and Economics 55, 148–167.

Jayaraman, S., 2008. Earnings volatility, cash flow volatility, and informed trading. Journal of Accounting Research 46, 809–851.

Karatzas, I., Shreve, S.E., 1988. Brownian Motion and Stochastic Calculus. Springer-Verlag, New York.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1336.

Lee, C.M., Ready, M.J., 1991. Inferring trade direction from intraday data. Journal of Finance 46, 733–746.

Li, H., Wang, J., Wu, C., He, Y., 2009. Are liquidity and information risks priced in the Treasury bond market? Journal of Finance 64, 467–503.

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

Madhavan, A., Richardson, M., Roomans, M., 1997. Why do security prices change? Review of Financial Studies 10, 1035–64.

Mohanram, P., Rajgopal, S., 2009. Is PIN priced risk? Journal of Accounting and Economics 47, 226–243.

Odders-White, E.R., Ready, M.J., 2008. The probability and magnitude of information events. Journal of Financial Economics 87, 227–248.

Revuz, D., Yor, M., 1991. Continuous Martingales and Brownian Motion. Springer-Verlag, Berlin.

Rogers, L.C.G., Williams, D., 2000. Diffusions, Markov Processes and Martingales: Vol. 2: Itô Calculus. 2nd ed., Cambridge University Press, Cambridge.

Rossi, S., Tinn, K., 2010. Man or machine? Rational trading without information about fundamentals. Working Paper.

Figure 1: Updating and Lambda in the Kyle Model.
This figure shows the probability (6) of an information event and Kyle's lambda (9) conditional on the order imbalance Y_t at date $t = 1/2$, when the unconditional probability of an information event is $\alpha = 0.3$, the standard deviation of liquidity trading is $\sigma = 0.3$, and the signal scale parameter is $\kappa = 0.005$. The range of the order imbalance in the plots is ± 3 times the standard deviation of Y_t.

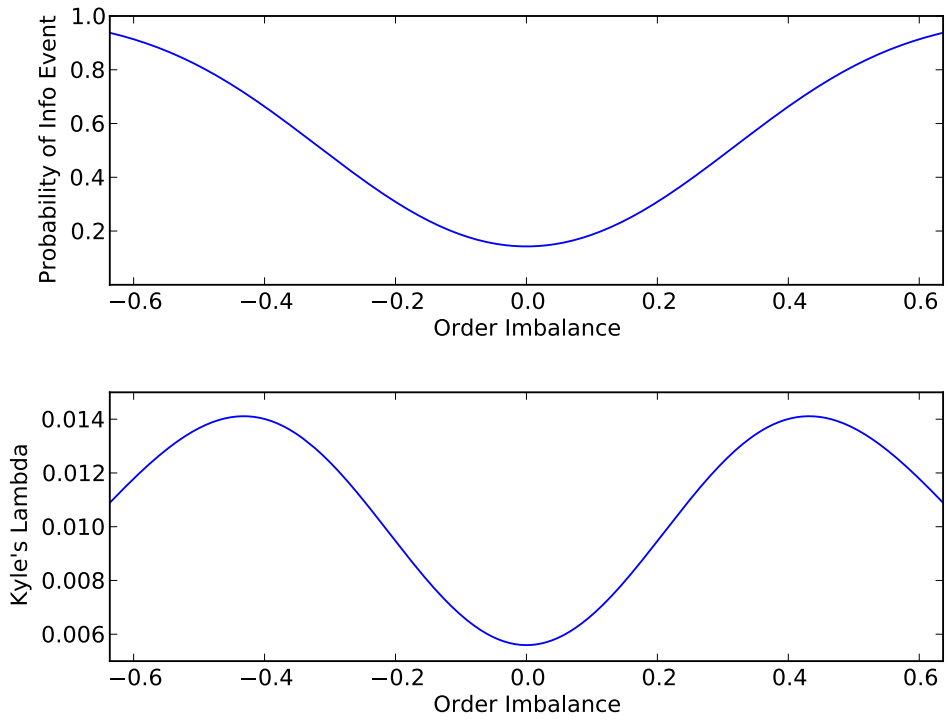

Figure 2: Monte Carlo with Daily Observations.
Estimates from 1000 simulated firm-months (22 trading days) using closing price/order flow observations only ($k = 0$). The model parameters are α = probability of an information event, κ = signal scale parameter, σ = standard deviation of liquidity trading, and δ = volatility of public information. The true parameters $\alpha = 30\%$, $\kappa = 0.5\%$, $\sigma = 30\%$, and $\delta = 1\%$ are denoted by the dashed vertical lines. Signal Standard Deviation is the unconditional signal standard deviation defined in (19). Its true value is 0.30%. Order Flow Component is the fraction of the return variance due to order flow information defined in (20). Its true value is 8.05%.

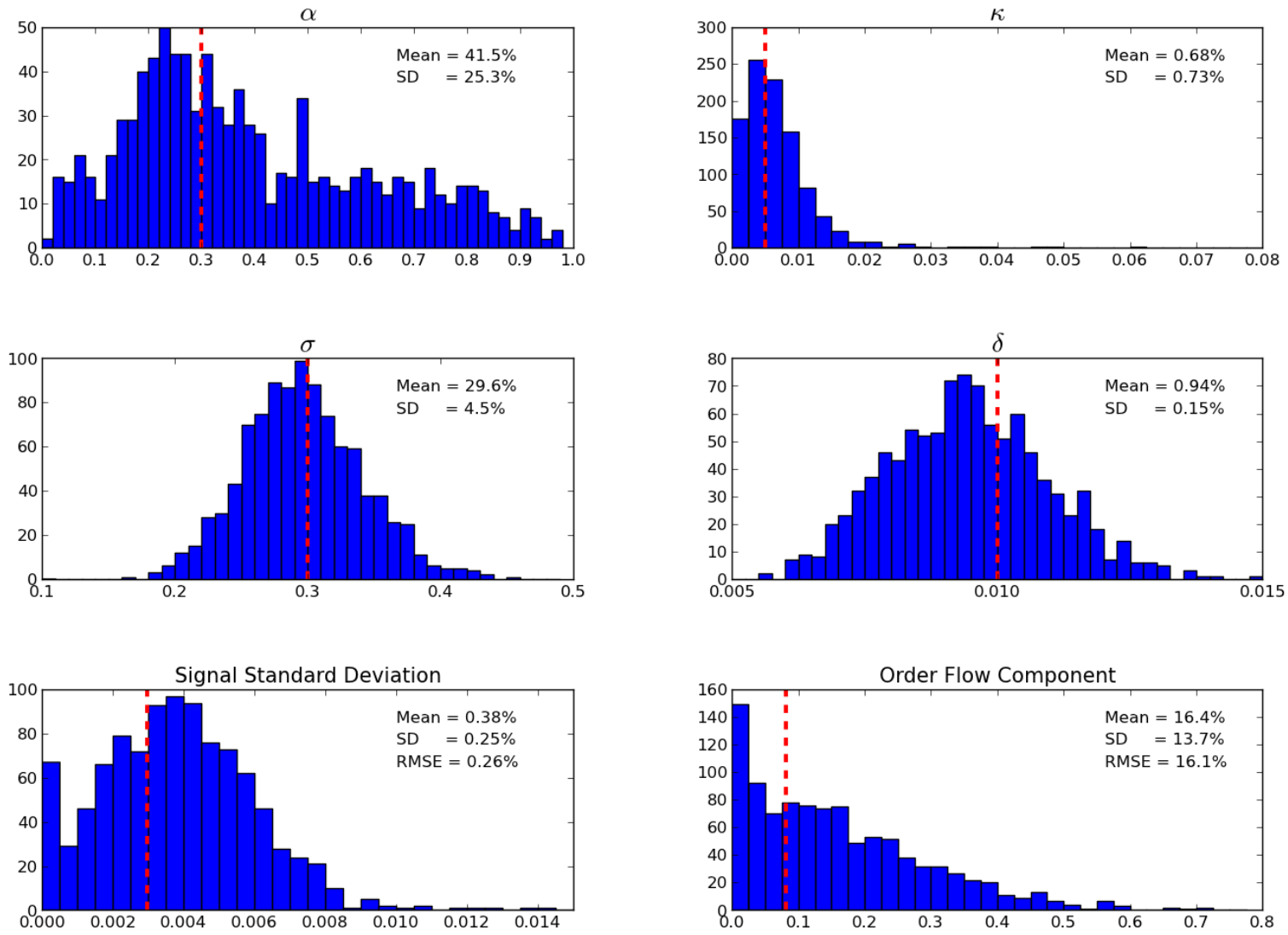

Figure 3: Monte Carlo with Intraday Observations.
Estimates from 1000 simulated firm-months (22 trading days) with hourly in addition to closing price/order flow observations ($k = 6$). The model parameters are α = probability of an information event, κ = signal scale parameter, σ = standard deviation of liquidity trading, and δ = volatility of public information. The true parameters $\alpha = 30\%$, $\kappa = 0.5\%$, $\sigma = 30\%$, and $\delta = 1\%$ are denoted by the dashed vertical lines. Signal Standard Deviation is the unconditional signal standard deviation defined in (19). Its true value is 0.30%. Order Flow Component is the fraction of the return variance due to order flow information defined in (20). Its true value is 8.05%.

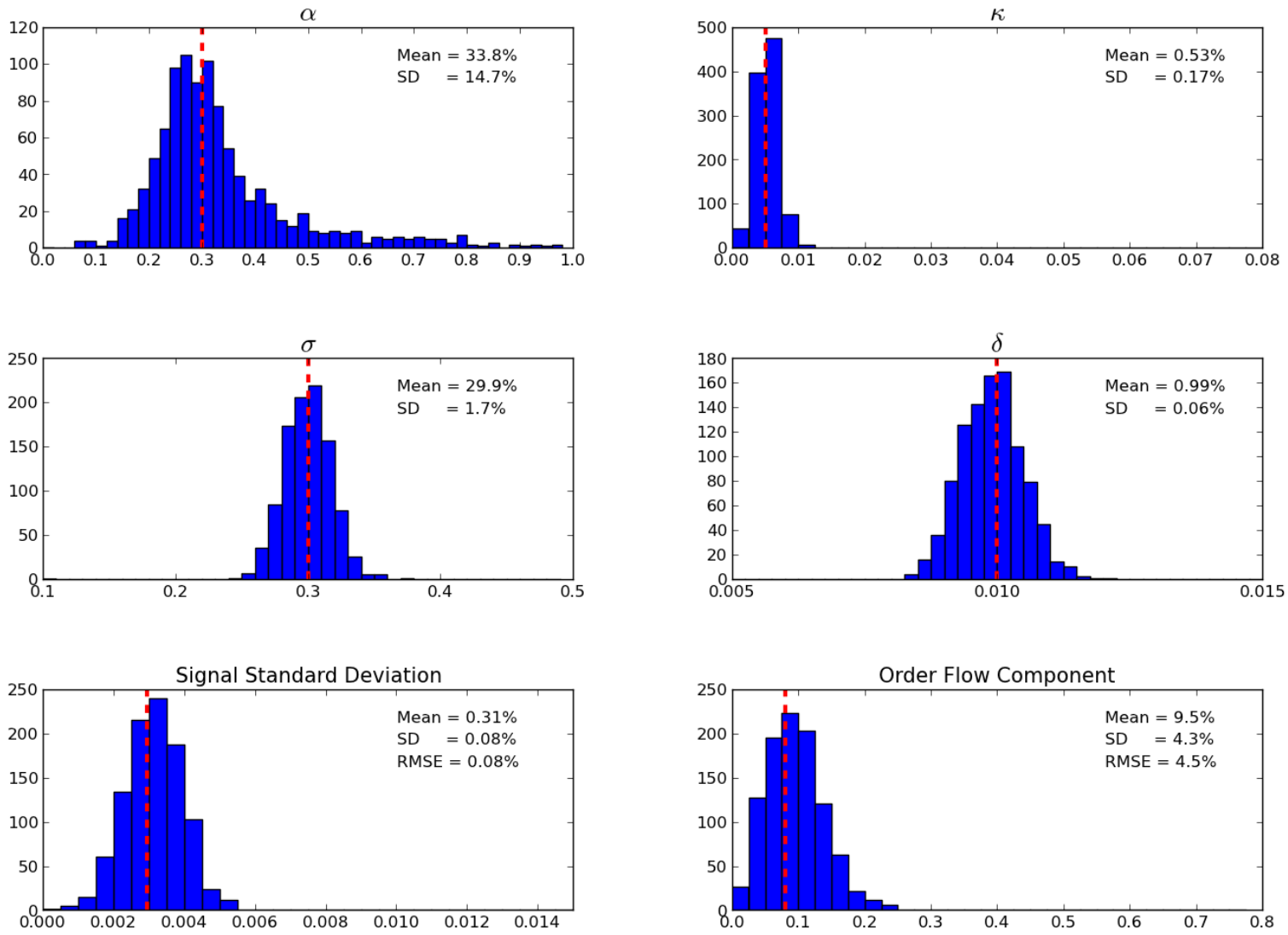

Figure 4: Monte Carlo with Lognormal Signal
Estimates from 1000 simulated firm-months (22 trading days) with hourly in addition to closing price/order flow observations ($k = 6$). The data are generated using a lognormal signal distribution with $W_i = \kappa_1 \exp(\kappa_2 \varepsilon_i - 0.5\kappa_2^2)$ with $\kappa_1 = 0.5\%$, $\kappa_2 = 0.25\%$, and standard normal ε_i. The model is estimated using the triangular signal distribution with parameters α = probability of an information event, κ = signal scale parameter, σ = standard deviation of liquidity trading, and δ = volatility of public information. The true parameters $\alpha = 30\%$, $\sigma = 30\%$, and $\delta = 1\%$ are denoted by the dashed vertical lines. Signal Standard Deviation is the unconditional signal standard deviation defined in (19). Its true value under the lognormal signal is 0.30%. Order Flow Component is the fraction of the return variance due to order flow information defined in (20). Its true value under the lognormal signal is 6.98%.

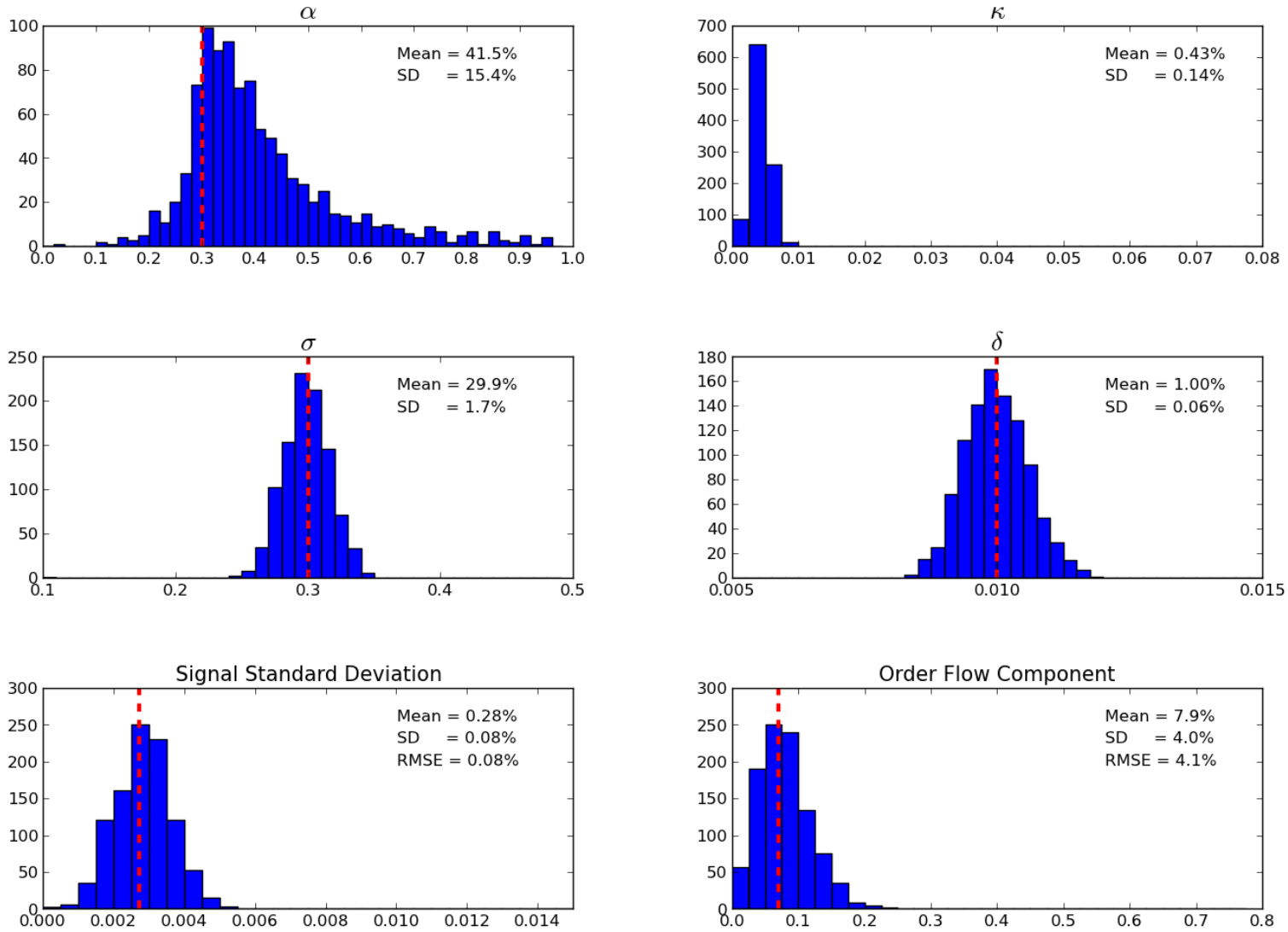

Figure 5: *Histograms of ASH Estimates.*
This figure plots the histograms of monthly maximum likelihood estimates for Ashland Inc. (ticker: ASH). The model is estimated from January 1993 through December 2012 using six hourly intraday bins ($k = 6$) in addition to closing prices and order imbalances. The model parameters are α = probability of an information event, κ = signal scale parameter, σ = standard deviation of liquidity trading, and δ = volatility of public information. Signal Standard Deviation is the unconditional signal standard deviation defined in (19), and Order Flow Component is the fraction of the return variance due to order flow information defined in (20).

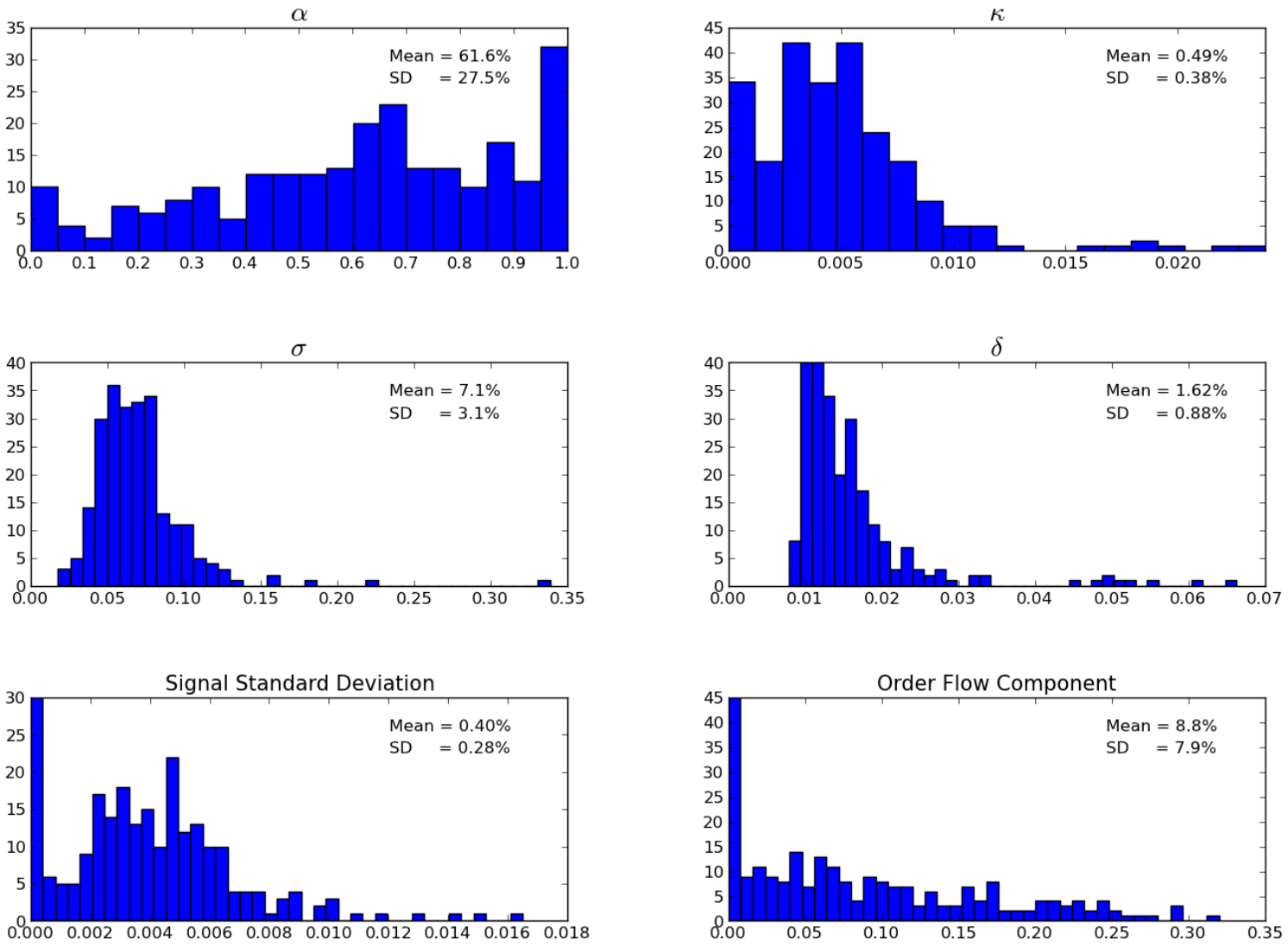

Figure 6: *Time Series of ASH Estimates.*
This figure plots the time series of monthly maximum likelihood estimates for Ashland Inc. (ticker: ASH). The model is estimated from January 1993 through December 2012 using six hourly intraday bins ($k = 6$) in addition to closing prices and order imbalances. The model parameters are α = probability of an information event, κ = signal scale parameter, σ = standard deviation of liquidity trading, and δ = volatility of public information. Signal Standard Deviation is the unconditional signal standard deviation defined in (19), and Order Flow Component is the fraction of the return variance due to order flow information defined in (20). The figure shows 90% confidence intervals estimated using the outer-product method (Hamilton, 1994). Standard errors for Signal Standard Deviation and Order Flow Component are estimated using the delta method.

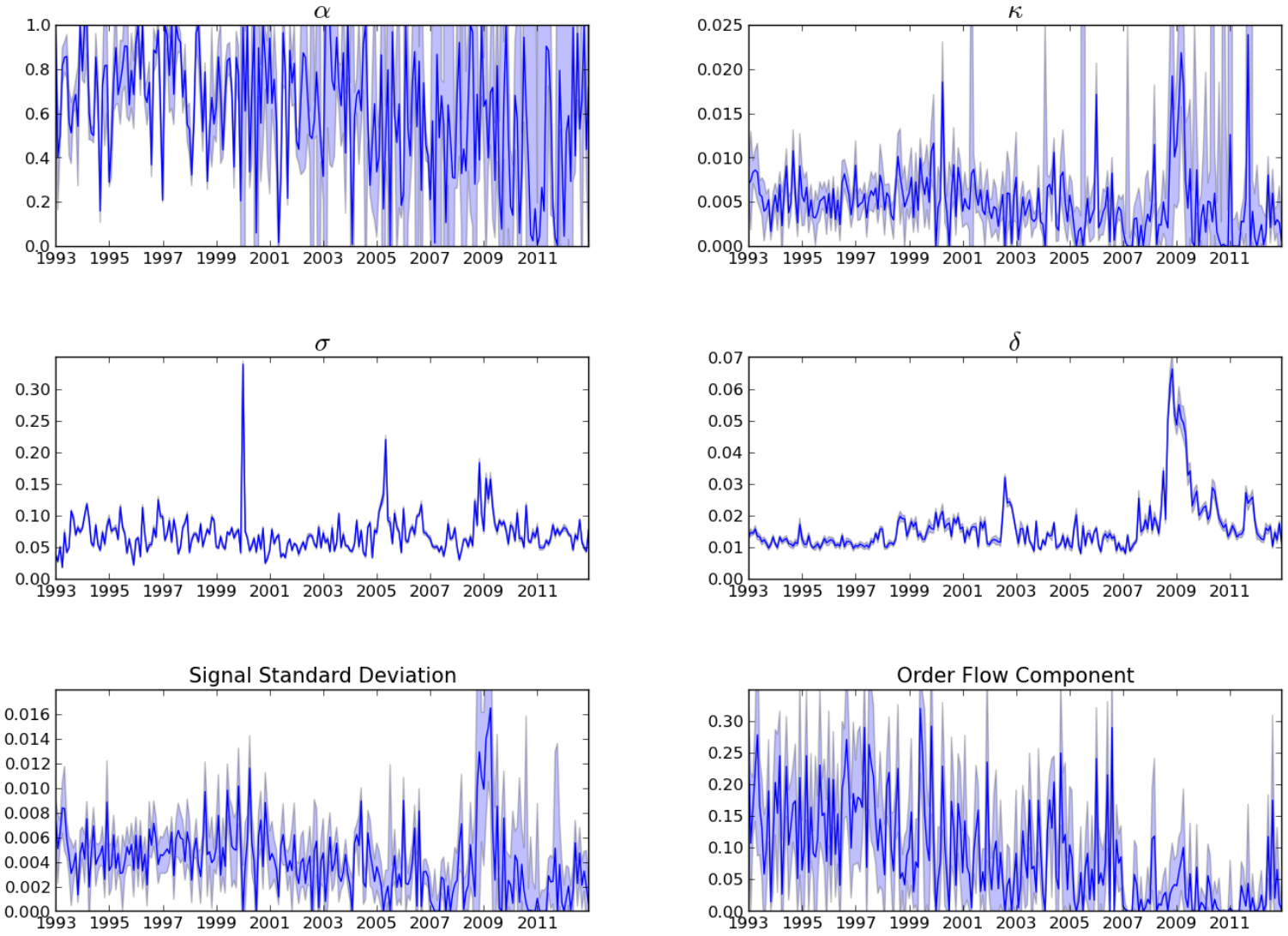

Table 1: Summary Statistics for ASH
This table presents summary statistics by odd years for daily returns, order imbalances, and volume for ASH. Returns are open-to-close returns, expressed in percentage points. Order imbalance is defined as shares bought less shares sold, expressed in units of one million shares. Volume is number of shares traded, expressed in units of one million shares.

Panel A. Returns

Year	Mean	SD	Min	Q1	Q2	Q3	Max
1993	0.03	1.27	-3.85	-0.61	0.00	0.70	4.45
1995	0.01	1.05	-2.33	-0.68	0.00	0.57	3.93
1997	0.00	1.18	-3.64	-0.77	0.00	0.67	4.26
1999	-0.06	1.36	-3.77	-1.02	-0.11	0.82	3.30
2001	0.11	1.44	-4.67	-0.66	0.11	1.01	5.64
2003	0.29	1.44	-4.35	-0.51	0.11	1.06	6.52
2005	-0.03	1.41	-6.46	-0.92	0.00	0.85	4.79
2007	-0.13	1.50	-4.96	-0.96	-0.17	0.74	7.91
2009	0.41	4.00	-13.40	-1.97	0.15	2.71	11.43
2011	-0.09	1.91	-6.28	-1.27	-0.08	1.10	5.66

Panel B. Order Imbalances

Year	Mean	SD	Min	Q1	Q2	Q3	Max
1993	0.010	0.072	-0.379	-0.016	0.006	0.035	0.303
1995	0.008	0.073	-0.404	-0.022	0.007	0.035	0.227
1997	0.017	0.080	-0.299	-0.028	0.008	0.055	0.299
1999	-0.001	0.067	-0.223	-0.032	-0.004	0.032	0.284
2001	0.030	0.064	-0.281	-0.002	0.028	0.062	0.221
2003	0.044	0.071	-0.357	0.012	0.038	0.076	0.322
2005	0.077	0.135	-0.261	0.004	0.052	0.126	1.030
2007	-0.012	0.088	-0.309	-0.055	-0.003	0.033	0.291
2009	-0.003	0.152	-0.612	-0.077	-0.000	0.069	0.450
2011	-0.013	0.082	-0.394	-0.057	-0.017	0.035	0.318

Panel C. Volume

Year	Mean	SD	Min	Q1	Q2	Q3	Max
1993	0.147	0.115	0.016	0.073	0.108	0.188	0.729
1995	0.174	0.119	0.025	0.093	0.141	0.226	0.800
1997	0.244	0.132	0.061	0.154	0.208	0.304	0.735
1999	0.220	0.124	0.047	0.137	0.194	0.266	0.919
2001	0.275	0.108	0.080	0.199	0.253	0.334	0.767
2003	0.366	0.189	0.124	0.239	0.316	0.458	1.391
2005	0.862	0.470	0.262	0.558	0.743	1.025	3.293
2007	0.723	0.302	0.280	0.508	0.651	0.879	2.426
2009	1.740	0.962	0.376	1.035	1.470	2.166	5.735
2011	0.927	0.530	0.309	0.613	0.786	1.067	3.671