Estimating the Probability of Informed Trading: A Bayesian approach

[1], Jim Griffin[1], Jaideep Oberoi*[2], and Samuel D. Oduro[3]

[1] *Department of Statistical Science, University College, Gower Street, London WC1E 6BT, United Kingdom.* Email: j.griffin@ucl.ac.uk

[2] *Kent Business School, University of Kent, Parkwood Road, Canterbury CT2 7FS, United Kingdom.* Tel: +44 1227 82 3865 Email: *j.s.oberoi@kent.ac.uk*

[3] *Data Science & Analytics, easyJet Airline Company, Hangar 89, London Luton Airport, Luton, Berdforshire, LU2 9PF, United Kingdom.* Email: *Samuel.DuaOduro@easyJet.com*

Abstract

The Probability of Informed Trading (PIN) is a widely used indicator of information asymmetry risk in the trading of securities. Its estimation using maximum likelihood algorithms has been shown to be problematic, resulting in biased or unavailable estimates, especially in the case of liquid and frequently traded assets. We provide an alternative approach to estimating PIN by means of a Bayesian method that addresses some of the shortcomings in the existing estimation strategies. The method leads to a natural quantification of the uncertainty of PIN estimates, which may prove helpful in their use and interpretation. We also provide an easy to use toolbox for estimating PIN.

JEL classification: C13, G12, G14

Keywords: PIN, software, Bayesian estimation, information asymmetry risk, robust estimation.

*Please send correspondence to j.s.oberoi@kent.ac.uk. Tel: +44 1227 82 3865

1. Introduction

The probability of informed trading (PIN) is a widely used measure of information asymmetry risk introduced in a sequence of papers following Easley and O'Hara (1992). In particular, the measures in Easley, Kiefer, O'Hara, and Paperman (1996) (EKOP-PIN for the initials of the authors) and Easley, Hvidkjaer, and O'Hara (2002) (EHO-PIN) have been frequently applied in studies related to illiquidity risk and informed trading. In this paper, we propose a Bayesian approach to estimating PIN in order to address well-documented problems with its estimation using maximum likelihood algorithms. We also provide the associated code in the form of a toolbox for use by researchers.

The importance of PIN is associated with its implications for trading costs, illiquidity risk and expected returns (see *e.g.* the theoretical models of Glosten and Milgrom, 1985; Kyle, 1985; Easley and O'Hara, 1987, 2004). It is based on the assumption that there are two types of agents that enter the market to trade: those wishing to exploit superior or private information (informed traders) and those that wish to trade for other reasons (variously referred to as uninformed, liquidity or noise traders). In theoretical models, market makers adjust their bid and ask quotes to reduce the risk of losses from trading against informed counterparties. As this may affect the expected returns on an asset, PIN has been used as the main explanatory variable or as a control in a large number of studies. These range from regressions related to asset pricing to event studies or other market analyses where private information may be involved. Consequently, the measure itself has undergone scrutiny and refinement over the years.

In the theoretical model, the key information observed by the market maker to estimate PIN is the order flow. As a result, PIN models specify a distribution for the signed trades (buys or sells) initiated by informed and uninformed traders and estimate the relevant parameters using the maximum likelihood estimator (MLE). The literature has found two practical drawbacks with the MLE. Firstly, the estimation algorithm leads too often to floating-point exceptions when the inputs to the likelihood function (numbers of buy and sell

trades) are large (see, *e.g.* Boehmer, Grammig, and Theissen, 2007; Jackson, 2013; Lei and Wu, 2005; Lin and Ke, 2011; Yan and Zhang, 2012). Secondly, the estimates of some of the underlying parameters of PIN (from existing MLE algorithms) often fall on the boundary of the parameter space. It is not clear whether the instances of boundary value solutions for the parameters arise due to the choice of initial values or due to model misspecification, but they could lead to bias and instability in PIN estimates.

Problems with the MLE algorithms have been acknowledged since the early work applying PIN. For instance, Easley and O'Hara (2004) introduce a modified factorization of the likelihood function to reduce the effects of such problems. Lin and Ke (2011) address the problems further by using an alternative factorization of the objective function. However, Yan and Zhang (2012) show that one still needs to choose initial values for the MLE maximizer carefully in order to achieve stable results. Thus the estimates are likely to be dependent on the choice of the initial values used by the optimizer, an issue of concern if the likelihood surface has several local maxima. In our applications, we observed that the parameter estimates were often very close to the starting values, suggesting the need for additional care. Other papers have further refined MLE estimation (see, *e.g.* Gan, Wei, and Johnstone, 2015), but the underlying issues remain. Our approach sidesteps these issues by using Bayesian estimation instead.

Several papers use PIN in cross-section or panel regressions (see, *e.g.* Chen, Goldstein, and Jiang, 2007; Christoffersen, Goyenko, Jacobs, and Karoui, 2018; Duarte and Young, 2009; Easley et al., 2002; Easley and O'Hara, 2004; Easley, Hvidkjaer, and O'Hara, 2010; Lai, Ng, and Zhang, 2014; Mohanram and Rajgopal, 2009; Vega, 2006, for just a few.). When PIN values cannot be calculated, the loss of observations in such regressions can be significant. Jackson (2013) reports that Easley et al. (2010) "lose firms representing nearly 24% of the market capitalization of the NYSE and AMEX" while in "Yan and Zhang (2012), the fraction of market capitalization lost grows from 2% in 1993 to 42% in 2004." Lost observations can lead to biased results, especially since larger firms (which have high

numbers of trades) are more likely to be affected.

Our paper contributes to the literature by addressing estimation problems identified in previous studies. We use Markov chain Monte Carlo (MCMC) methods to explore the entire posterior distribution of model parameters in a Bayesian analysis, thereby avoiding the numerical instability problem faced by MLE maximizers.[1] Specifically, we propose a direct prior for the PIN and then implement a Gibbs sampling algorithm to estimate the model. The Bayesian approach also provides a natural way of quantifying uncertainty in point estimates of the PIN (using credible intervals) from its posterior distribution. Most available methods and the majority of papers in the literature do not pay much attention to the uncertainty of PIN estimates. Given the potential for model misspecification (see Gan, Wei, and Johnstone, 2017), it may be useful for researchers to be aware of the uncertainty surrounding their point estimates of PIN. In addition, our approach leads to reliable estimation of PIN at daily frequency using only 26 intraday observations, as compared to the usual recommendation in the literature for a minimum of 60 daily observations resulting in quarterly estimates. Higher frequency estimates offer opportunities to use PIN for more studies, for instance looking at changes in the measure surrounding corporate event announcements or actions.

The improvement in estimation proposed here matters not simply for its own sake. One of the debates about PIN is that it does not represent the probability that it purports to measure (see, *e.g.* Aktas, De Bodt, Declerck, and Van Oppens, 2007; Duarte and Young, 2009). However, if the bias in PIN is caused by issues such as numerical problems, then the debate cannot be fully resolved, because the bias will not be well-understood.

It is important to note that PIN over any period is estimated independently of its values in other periods. Corner solutions and local optima in the chosen optimizer would introduce instability of a misleading nature, for instance caused simply by a change in volume. In turn, this would be unhelpful for inference in studies that use PIN for a sequence of dates or periods. The potential to avoid such sources of instability opens up opportunities to use the measure

[1] A referee has pointed out a way to understand why the Bayesian procedure avoids corner solutions is that it involves integration rather than maximization of the posterior.

in more applications. PIN is usually calculated with the daily numbers of buy and sell trades over a period of between one quarter and one year. Given the speed of markets, it would be useful to have, say, a time series of daily estimates of PIN by assuming that news arrives and is absorbed by markets at shorter intervals than one day. Standard estimation methods over shorter horizons are considered problematic partly because of the instability of the estimates, so most studies have been limited to lower frequency cross-sectional applications of PIN. A notable exception is that of Brennan, Huh, and Subrahmanyam (2018), who used PIN in an event study setting. In order to achieve this, they first estimate PIN over a longer period (two months) and then update the estimates using Bayesian updating on a daily basis. Our approach is more direct and can work over any of the frequencies. In this paper, we demonstrate how our methodology can be used to produce daily estimates of PIN and its credible intervals over twelve years for five stocks. We also provide quarterly estimates over the same period, and in both cases, compare the estimates to those obtained by MLE.

We also conduct a simulation exercise to show that the Bayesian method performs competitively with respect to the MLE, even when the data is simulated from the underlying model. The caveat to the simulation study is that the observed trades data is known to be much more challenging than that generated by the underlying model. For instance, Venter and De Jongh (2006) and Gan et al. (2017) specifically show that the clusters of buy and sell trades generated by the EHO-PIN model are more differentiated and have a lower scale compared to the observed data. An additional simulation reported in the Appendix considers a case where the data is generated from a modified model, making the PIN misspecified by definition. Although this is not the focus of the paper, the exercise may also help support the notion that the PIN model, even when biased, can be informative in the contexts in which it is widely applied.

The paper is organized as follows. The next section provides background in the form of a brief description of the EHO-PIN model along with the MLE procedure. Section 3 details the Bayesian procedure for estimation. In Section 4 we demonstrate our results using data

on five stocks over a twelve year period. Although the ordering is not customary, we then present simulation results in Section 5, as these results are meant to support our analysis rather than provide the main illustration. We then briefly conclude.

2. Background

This section provides background on the EHO version of the PIN model and its estimation under the MLE approach.

2.1. The EHO-PIN Model

In the theoretical market microstructure setting (see *e.g.* Glosten and Milgrom, 1985), informed traders act on advantage while uninformed traders buy or sell for reasons other than the possession of superior knowledge of the fundamental value of the asset. In such a setting, Easley et al. (1996) and Easley, Kiefer, and O'Hara (1997) estimated models based on the information structure in Easley and O'Hara (1992). The main idea behind the model is that an unusual imbalance between buy and sell trades reflects the activity of informed traders.

The model described by Easley et al. (2002) assumes that within any trading day, the number of buyer and seller initiated trades from informed and uninformed traders are realizations of independent Poisson distributions whose mean depends on whether no news, good news or bad news occurs on that day (a representation of the model as a probability tree is provided in Figure 1). We will think about the natural generalisation of the model where the type of news is fixed over intervals at other frequencies, for example over a $15-$minute interval. As a result, we will substitute the phrase *trading period* instead of trading day in this paper. The model assumes that the probability of news (good or bad) in any trading period is α. Given that there is news, the probability that an asset value will be negatively affected by the news event is δ (which implies that the probability of a positive effect is

$1-\delta$). In any given trading period, liquidity traders are present in the market to either buy or sell the asset. In a bad news period, informed traders expect an adverse effect on the value of the asset, and are therefore likely to sell the asset. The order arrivals in a bad news period are assumed to follow independent Poisson distributions with means μ for the informed traders, and λ_b and λ_s for liquidity buy and sell traders respectively. This implies that the total numbers of buyer and seller initiated trades in a bad news period are Poisson distributed with means λ_b and $\lambda_s+\mu$ respectively. Similarly, in a good news period, the total numbers of buy and sell trades are Poisson distributed with means $\lambda_b+\mu$ and λ_s respectively. Finally, in a no news period, informed traders will not participate and the total numbers of buys and sells are Poisson distributed with means λ_b and λ_s respectively. In practice, we do not observe the arrival of traders or the occurrence of a news event and these must be inferred from the observable trade data.

[Insert Figure 1 near here]

2.2. MLE Approach

Let B_t and S_t be the total numbers of buy and sell trades in trading period t, respectively. The joint likelihood function for a sample of $t = 1, \ldots, T$, trading periods is given as follows

$$\begin{aligned} L\left(\Theta|B,S\right) &= \prod_{t=1}^{T}\left[\alpha\delta\frac{e^{-(\mu+\lambda_s)}\left(\mu+\lambda_s\right)^{S_t}}{S_t!}\frac{e^{-\lambda_b}\left(\lambda_b\right)^{B_t}}{B_t!}+\alpha\left(1-\delta\right)\frac{e^{-(\mu+\lambda_b)}\left(\mu+\lambda_b\right)^{B_t}}{B_t!}\frac{e^{-\lambda_s}\left(\lambda_s\right)^{S_t}}{S_t!}\right. \\ &\quad \left.+\left(1-\alpha\right)\frac{e^{-\lambda_b}\left(\lambda_b\right)^{B_t}}{B_t!}\frac{e^{-\lambda_s}\left(\lambda_s\right)^{S_t}}{S_t!}\right], \end{aligned} \tag{1}$$

where $\Theta = (\alpha, \delta, \mu, \lambda_b, \lambda_s)$, $B = (B_1, \ldots, B_T)$ and $S = (S_1, \ldots, S_T)$. Easley et al. (2002) estimate the vector of parameters Θ by maximizing equation 1. In this model the PIN is defined as

$$PIN = \frac{\alpha\mu}{\alpha\mu+\lambda_s+\lambda_b}. \tag{2}$$

This is the ratio of expected informed trading to expected total trades. When we do not observe whether a trade was initiated by a buy order or a sell order, classification of trades as buys and sells can be carried out using the information available, for instance by using a rule based on the Lee and Ready (1991) trade classification algorithm.

However, MLE has been shown to lead to biased samples. In some cases, this results simply from the size of the numbers that enter the estimator, leading to "NaN" error codes. As noted by Lin and Ke (2011) and Yan and Zhang (2012), the daily number of trades has grown large. Therefore it is possible for the likelihood function to produce a number larger (or smaller) than the largest (smallest) acceptable value of a computer software. However, we also find that smaller numbers of trades (*e.g.*, when counted at 15−minute intervals) still lead to less stable estimates than the Bayesian method, when plotted over time. This may be related to the fact that there are only 26 trading periods of length 15 minutes in a day.

3. The Bayesian Estimation Approach

Our goal is to learn about PIN and its underlying parameters from observed transaction data. The MLE approach assumes that the model parameters are unknown but fixed. In Bayesian inference, we express the uncertainty about the unknown model parameters through the rules of probability. We achieve this through Bayes' rule which states that the probability of the parameter set Θ given the observed data is

$$\begin{aligned} p(\Theta|B,S) &= \frac{p(B,S,\Theta)}{p(B,S)} \\ &= \frac{p(\Theta)\,p(B,S|\Theta)}{p(B,S)} \\ &\propto p(\Theta)\,p(B,S|\Theta). \end{aligned} \tag{3}$$

The denominator in Equation 3, $p(B,S) = \int p(\Theta)\,p(B,S|\Theta)d\Theta$, is a normalizing constant. It guarantees that $p(\Theta|B,S)$ is a well defined probability density function. The term $p(\Theta)$,

referred to as the prior density, is not dependent on the data. It is used to express the prior knowledge and uncertainty about the model parameters before observing the data. The term $p(B, S|\Theta)$, usually referred to as the likelihood function is the probability density function of the data conditional on the model parameters. In Bayesian inference, the primary object of interest is $p(\Theta|B, S)$, which is referred to as the posterior density. From the posterior density, we can compute point estimates like the mean and mode as well as credible intervals for the model parameters. We employ MCMC methods to infer the parameters of the EHO model. These MCMC methods explore the entire support of the posterior distribution of the model parameters. In what follows we provide a description of the MCMC methods.

3.1. The Gibbs Sampler

The Gibbs Sampler is an MCMC algorithm which generates a sample from posterior distributions whose kernels are known standard probability density functions. The algorithm uses the full conditional distribution of the posterior distribution. If the parameters are $\theta_1, \ldots, \theta_k$ then the full conditional distribution for θ_j is $p(\theta_j|\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k, y)$. The algorithm proceeds by updating each parameter (or block of parameters) in turn from its full conditional distribution by sampling a value from its full conditional distribution (with all other parameters set to their current values). Each iteration of the Gibbs sampler involves updating all parameters. A summary of the Gibbs sampler algorithm is as follows:

- Step 0 : Initialize $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}$
- Step 1 : Draw once from $p(\theta_1|\theta_2^{(0)}, \ldots, \theta_k^{(0)}, y)$ to obtain $\theta_1^{(1)}$
- Step 2 : Draw once from $p(\theta_2|\theta_1^{(1)}, \ldots, \theta_k^{(0)}, y)$ to obtain $\theta_2^{(1)}$
- ...
- Step k : Draw once from $p(\theta_k|\theta_1^{(1)}, \ldots, \theta_{k-1}^{(1)}, y)$ to obtain $\theta_k^{(1)}$
- Repeat Steps 1 to k for say G times to generate G Monte Carlo draws from the full conditionals.

The posterior expectation of any function of the parameters can be estimated using the output from the Gibbs sampler using the Monte Carlo average

$$\mathrm{E}[f(\theta_1,\ldots,\theta_k)|y] = \frac{1}{G-G_0}\sum_{i=G_0+1}^{G} f\left(\theta_1^{(i)},\ldots,\theta_k^{(i)}\right)$$

where G_0 is a burn-in time that is used to remove the dependence of the $\theta_1^{(i)},\ldots,\theta_k^{(i)}$ on the initial values.

3.2. Joint Density Of Buy And Sell Orders

As per the EHO-PIN model, all trading periods can be classified into three types. Let this classification be labelled D_t, such that

$$D_t = \begin{cases} 1, & \text{bad news, with probability} \quad \omega_1 = \alpha\delta \\ 2, & \text{good news, with probability} \quad \omega_2 = \alpha(1-\delta) \\ 3, & \text{no news, with probability} \quad \omega_3 = 1-\alpha. \end{cases}$$

where ω_D is the probability of news type D. The model assumes that, when $D_t = 1$, informed traders take a short position and liquidity traders either buy or sell the asset for reasons other than information. Since only liquidity traders make buy trades, the total number of buy trades (B_t) is Poisson distributed with mean λ_b. The numbers of sell trades by informed traders (S_t^i) and by liquidity traders (S_t^u) follow independent Poisson distributions with means μ and λ_s respectively. Thus the total number of sell trades ($S_t = S_t^i + S_t^u$) follows a Poisson distribution with mean $\mu + \lambda_s$. Using a similar reasoning for each type of news event, we can state conditional distributions of the numbers of buy and sell trades as

$$S_t|D_t = 1 \sim Pn\left(\mu+\lambda_s\right) \qquad S_t|D_t = 2 \sim Pn\left(\lambda_s\right) \qquad S_t|D_t = 3 \sim Pn\left(\lambda_s\right)$$
$$B_t|D_t = 1 \sim Pn\left(\lambda_b\right) \qquad B_t|D_t = 2 \sim Pn\left(\mu+\lambda_b\right) \qquad B_t|D_t = 3 \sim Pn\left(\lambda_b\right),$$

where $Pn(.)$ is the probability mass function of a Poisson random variable.[2]

The underlying process D_t is not observable but can be inferred from transaction data, as a missing data problem within the Bayesian framework. Since we do not observe a bad, good or no news period as well as the arrival of liquidity and informed traders, we employ the data augmentation procedure to impute these missing observations. We do this by directly sampling from the posterior distribution of D_t conditional on the available data. For a detailed review of data augmentation, see Van Dyk and Meng (2001).

The joint density of buy and sell orders using the data augmentation procedure is

$$\begin{aligned} P(B_t, S_t | D_t, \Theta) &= \left[f_1(B_t, S_t, \Theta) \right]^{d_{t,1}} \left[f_2(B_t, S_t, \Theta) \right]^{d_{t,2}} \left[f_3(B_t, S_t, \Theta) \right]^{d_{t,3}} \\ &= \left[\frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!} \frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!} \frac{e^{-\lambda_s}\lambda_s^{S_t - S_t^i}}{(S_t - S_t^i)!} \right]^{d_{t,1}} \left[\frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!} \frac{e^{-\lambda_s}\lambda_s^{S_t}}{S_t!} \frac{e^{-\lambda_b}\lambda_b^{B_t - B_t^i}}{(B_t - B_t^i)!} \right]^{d_{t,2}} \\ &\times \left[\frac{e^{-\lambda_s}(\lambda_s)^{S_t}}{S_t!} \frac{e^{-\lambda_b}(\lambda_b)^{B_t}}{B_t!} \right]^{d_{t,3}}, \end{aligned}$$

where $d_{t,j} = \mathbf{1}_{\{D_t = j\}}$, for $j = 1, 2, 3$. This is obtained by combining the likelihood functions for the joint densities of buy and sell orders under each of the realizations of D_t. The derivations of these joint densities are provided in Appendix B.1.

3.3. Prior Distributions And MCMC Sampler

Since we employ a Bayesian approach to estimating the parameters of the PIN model, it is important to choose appropriate prior distributions for the parameters and write down the posterior distribution.

[2] $Pn(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}$

3.3.1. Prior distributions

We specify a prior for each of the parameters of the model. The parameters controlling the mean numbers of informed and uninformed trades are given a hierarchical prior

$$\lambda_s \sim \mathcal{G}a(a,\eta), \quad \lambda_b \sim \mathcal{G}a(b,\eta), \quad \mu \sim \mathcal{G}a(a,\eta).$$

where $\mathcal{G}a(.)$ denotes the gamma probability density function. This prior implies that prior number of buys (sells) on a good (bad) news day follows a $\mathcal{G}a(a+b,\eta)$ distribution and the prior number of buy and sells on a no news data follow $\mathcal{G}a(b,\eta)$. The proportion of informed trades on a news day has a beta distribution with parameters a and b. The hyperparameters a and b can be chosen to tune this prior distribution. The parameter η controls the overall level of trading and is given a vague, proper prior $\eta \sim \mathcal{G}a(0.001, 0.001)$. The parameters α and δ are both given uniform distributions on the interval $(0,1)$.

3.3.2. The Markov chain Monte Carlo Sampler

The use of conjugate priors allows the Gibbs Sampler to be easily applied to sample from the posterior distribution. From Bayes' theorem, the posterior density for the parameter set $\Theta = (\alpha, \delta, \mu, \lambda_s, \lambda_b, \eta)$ and the classification indicators $D = (D_1, \ldots, D_T)$ is proportional to the product of the likelihood and prior. If we denote T_1, T_2 and T_3 as the number of periods with bad, good and no news arrivals, then the posterior density can be written as

$$\begin{aligned}
P(\Theta, D|B,S) &\propto P(\Theta) \prod_{t=1}^{T} \left[P(B_t, S_t|D_t, \Theta) P(D_t|\Theta) \right] \\
&= \eta^{2a+b} \mu^{b-1} e^{-\eta\mu} \lambda_s^{a-1} e^{-\eta\lambda_s} \lambda_b^{a-1} e^{-\eta\lambda_b} \eta^{0.001-1} \exp\{-0.001\eta\} \alpha^1 (1-\alpha)^1 \delta^1 (1-\delta)^1 \\
&\times \left[(\alpha\delta)^{T_1} (\alpha(1-\delta))^{T_2} (1-\alpha)^{T_3} \right] \prod_{t=1}^{T} \left[\frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!} \frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!} \frac{e^{-\lambda_s}\lambda_s^{S_t - S_t^i}}{(S_t - S_t^i)!} \right]^{d_{t,1}} \\
&\times \left[\frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!} \frac{e^{-\lambda_s}\lambda_s^{S_t}}{S_t!} \frac{e^{-\lambda_b}\lambda_b^{B_t - B_t^i}}{(B_t - B_t^i)!} \right]^{d_{t,2}} \left[\frac{e^{-\lambda_s}(\lambda_s)^{S_t}}{S_t!} \frac{e^{-\lambda_b}(\lambda_b)^{B_t}}{B_t!} \right]^{d_{t,3}}. \qquad (4)
\end{aligned}$$

The full conditional distributions of the parameters of interest are

$$\alpha \sim \mathcal{B}e\left(1 + T_1 + T_2,\ \ 1 + T_3\right), \tag{5a}$$

$$\delta \sim \mathcal{B}e\left(1 + T_1,\ \ 1 + T_2\right), \tag{5b}$$

$$\mu \sim \mathcal{G}a\left(b + \sum_{t=1}^{T}[(S_t^i)^{d_{t,1}} + (B_t^i)^{d_{t,2}}],\ \eta + T_1 + T_2\right), \tag{5c}$$

$$\lambda_s \sim \mathcal{G}a\left(a + \sum_{t=1}^{T}[S_t^{d_{t,1}} - (S_t^i)^{d_{t,1}} + S_t^{d_{t,2}} + S_t^{d_{t,3}}],\quad \eta + T_1 + T_2 + T_3\right), \tag{5d}$$

$$\lambda_b \sim \mathcal{G}a\left(a + \sum_{t=1}^{T}[B_t^{d_{t,1}} + B_t^{d_t,2} - (B_t^i)^{d_{t,2}} + B_t^{d_{t,3}}],\quad \eta + T_1 + T_2 + T_3\right), \tag{5e}$$

$$\eta \sim \mathcal{G}a(0.001 + 2a + b, 0.001 + \mu + \lambda_s + \lambda_b) \tag{5f}$$

where $T = T_1 + T_2 + T_3$ is the total number of trading periods in the sample. In appendix B.2, we provide the derivation of the full conditional distributions.

The steps of the MCMC sampler to estimate the parameters in our PIN model are given below.

- Start with classification $\mathbf{D}^{(0)}$ of (B_t, S_t)
- Initialize the parameters $\Theta^{(0)}$=$\left(\alpha^{(0)}, \delta^{(0)}, \lambda_s^{(0)}, \lambda_b^{(0)}, \mu^{(0)}, \eta^{(0)}\right)$
- Repeat for $k = 1$ to G sweeps
  - Repeat for $t = 1$ to T
    - If $D_t = 1$, update $B_t^{i(k)} \sim Binomial(B_t, \frac{\mu^{(k-1)}}{\mu^{(k-1)}+\lambda_b^{(k-1)}})$
    - If $D_t = 2$, update $S_t^{i(k)} \sim Binomial(S_t, \frac{\mu^{(k-1)}}{\mu^{(k-1)}+\lambda_s^{(k-1)}})$
  - Update $\mu^{(k)}|\lambda_s^{(k-1)}, \lambda_b^{(k-1)}, \eta^{(k-1)}, \alpha^{(k-1)}, \delta^{(k-1)}, B_t^{i(k)}, S_t^{i(k)}$
  - Update $\lambda_s^{(k)}|\lambda_b^{(k-1)}, \eta^{(k-1)}, \alpha^{(k-1)}, \delta^{(k-1)}, B_t^{i(k)}, S_t^{i(k)}, \mu^{(k)}$
  - Update $\lambda_b^{(k)}|\eta^{(k-1)}, \alpha^{(k-1)}, \delta^{(k-1)}, B_t^{i(k)}, S_t^{i(k)}, \mu^{(k)}, \lambda_s^{(k)}$
  - Update $\eta^{(k)}|\alpha^{(k-1)}, \delta^{(k-1)}, B_t^{i(k)}, S_t^{i(k)}, \mu^{(k)}, \lambda_s^{(k)}, \lambda_b^{(k)}$
  - Update $\alpha^{(k)}|\delta^{(k-1)}, B_t^{i(k)}, S_t^{i(k)}, \mu^{(k)}, \lambda_s^{(k)}, \lambda_b^{(k)}, \eta^{(k)}$

- Update $\delta^{(k)}|B_t^{i(k)}, S_t^{i(k)}, \mu^{(k)}, \lambda_s^{(k)}, \lambda_b^{(k)}, \eta^{(k)}, \alpha^{(k)}$
- Compute $L_1 = \log \omega_1^{(k)} - \left(\mu^{(k)} + \lambda_s^{(k)} + \lambda_b^{(k)}\right) + B_t \log \lambda_b^{(k)} + S_t \log \left(\lambda_s^{(k)} + \mu^{(k)}\right)$
- Compute $L_2 = \log \omega_2^{(k)} - \left(\mu^{(k)} + \lambda_s^{(k)} + \lambda_b^{(k)}\right) + S_t \log \lambda_s^{(k)} + B_t \log \left(\lambda_b^{(k)} + \mu^{(k)}\right)$
- Compute $L_3 = \log \omega_3^{(k)} - \left(\lambda_s^{(k)} + \lambda_b^{(k)}\right) + S_t \log \lambda_s^{(k)} + B_t \log \lambda_b^{(k)}$
- compute $\chi = \max(L_1, L_2, L_3)$
- Compute $p_1 = \frac{e^{L_1 - \chi}}{\sum_{j=1}^{3} e^{L_j - \chi}}$, $p_2 = \frac{e^{L_2 - \chi}}{\sum_{j=1}^{3} e^{L_j - \chi}}$ and $p_3 = \frac{e^{L_3 - \chi}}{\sum_{j=1}^{3} e^{L_j - \chi}}$
- Update $\mathbf{D_t}^{(k)}$, the classification of (B_t, S_t) by sampling from the multinomial distribution with probability (p_1, p_2, p_3),

where p_1, p_2 and p_3 are the probabilities that at the beginning of the trading period there will be bad news, good news and no news respectively.

The algorithm yields G random samples drawn from the posterior distributions of the parameters α, δ, μ, λ_s, λ_b, and of PIN. Discarding say G_0 initial draws of each posterior sample and taking the average of the remaining, we obtain the estimated central value of each parameter. As we also have the posterior distributions of the parameters and of PIN, we can easily view the uncertainty around the PIN estimate as well.

In Appendix A, we provide a description of the BayesPIN toolbox for MATLAB that accompanies this paper online. This toolbox implements the MCMC algorithm for a given input of numbers of buy and sell trades.

We next demonstrate the procedure on real data and discuss the advantages of our approach.

4. Empirical Illustration

In order to demonstrate the methodology, we estimate PIN for five stocks on a daily basis over the period 5th January 2004 to 31st December 2015. We also estimate the quarterly PIN using daily counts of buy and sell trades. To compare the estimates with those from an MLE package, we use the *InfoTrad* package available in R (Çelik and Tiniç, 2018).

4.1. Data

The five stocks are IBM, Coca Cola (KO), Boeing (BA), Walt Disney (DIS), and Exxon Mobil Corporation (XOM), all large highly-traded stocks, but from different industries. We use millisecond time stamped National Best Bid Offer (NBBO) quotes and trades from TickData, a high frequency data vendor with expertise on generating NBBO data. Holden and Jacobsen (2014) argue that NBBO data is likely to contain fewer errors compared with raw quotes from individual exchanges. The data cover a total of $3,020$ trading days.

In terms of data cleaning procedures, we follow Korajczyk and Sadka (2008) by excluding all transactions that occurred outside the normal trading hours as well as weekend trades. We also remove all transactions that had negative prices as well as negative prevailing spreads. We further exclude all cases where the transaction price was higher (lower) than the ask (bid) price by more than 50 times the tick size ($0.01), or 50 cents.

The required input for estimating PIN is a sequence of numbers of buy and sell transactions at the chosen trading period frequency. We use the Lee and Ready (1991) algorithm to classify the individual transactions into buyer and seller initiated trades, which we then aggregate over $15-$minute intervals or daily intervals as required.

In Table 1, we provide a summary of the number of buyer and seller initiated trades for the assets over the two intervals. It can be observed that the daily buyer and seller initiated trades for these assets are large enough to potentially cause floating point exceptions. Corresponding scatter plots of numbers of buy and sell trades are shown in Figure 2 for the

daily frequency and in Figure 3 for the 15−minute frequency. The median number of buys exceeds the median number of sells for all of the stocks. The median number of buy trades ranges between 12,199 and 35,894 across the 5 stocks at the daily level, and between 380 and 1,165 at the 15-minute frequency.

Asset	Sampling Freq.		Min	Median	Mean	Max
BA	Daily	Buys	785	12,199	12,717	109,214
		Sells	646	11,221	11,806	103,212
	15 minutes	Buys	14	380	491	15,866
		Sells	1	351	456	16,091
DIS	Daily	Buys	1,086	19,185	20,195	186,073
		Sells	980	16,543	17,432	138,201
	15 minutes	Buys	12	607	780	31,128
		Sells	7	522	673	28,220
IBM	Daily	Buys	1,582	12,953	14,043	76,662
		Sells	1,510	12,837	13,826	66,627
	15 minutes	Buys	3	410	542	11,427
		Sells	7	407	534	12,478
KO	Daily	Buys	1,343	20,375	20,675	112,876
		Sells	1,298	17,766	17,925	117,509
	15 minutes	Buys	20	636	798	19,397
		Sells	22	549	692	18,277
XOM	Daily	Buys	1,716	35,894	39,807	260,535
		Sells	1,593	31,283	35,640	245,164
	15 minutes	Buys	27	1,165	1,537	27,425
		Sells	12	1,025	1,376	24,066

Table 1: Numbers of buy and sell trades counted at two frequencies

[Insert Figures 2 and 3 near here]

In our data set there are transactions on each trading day, though when sampling at 15−minute intervals there are some periods in which either there is no buy or no sell (but not both). In order for the algorithm to work, there should be a minimum of 1 trade (buy or sell) in every trading period.

4.2. Daily PIN estimation results

For each day, we estimate the PIN using trading periods of 15 minutes. We summarize and plot both the MLE benchmark and the Bayesian estimates below. For the Gibbs sampler, we chose the number of sweeps (G) to be $25,000$ with a burn-in (G_0) of $5,000$.

As a benchmark, we provide the MLE estimates using the R package *InfoTrad* (see Çelik and Tiniç, 2018) in Table 2. We use the Yan and Zhang (2012) grid search approach with likelihood function factorization proposed by Lin and Ke (2011), which are accepted in the literature so far to perform best on empirical data. The corresponding results for the Bayesian approach are provided in Table 3.

Parameter	Statistic	BA	DIS	IBM	KO	XOM
α	Min	0.038	0.038	0.038	0.038	0.038
	Median	0.269	0.308	0.269	0.269	0.286
	Mean	0.282	0.293	0.280	0.289	0.292
	Max	0.810	0.731	0.808	0.836	0.846
δ	Min	0.000	0.000	0.000	0.000	0.000
	Median	0.250	0.250	0.235	0.250	0.000
	Mean	0.379	0.364	0.377	0.373	0.297
	Max	1.000	1.000	1.000	1.000	1.000
μ	Min	29.1	31.7	38.2	27.2	53.2
	Median	498.6	704.2	552.8	769.9	1328.7
	Mean	596.9	876.3	667.4	942.0	1614.8
	Max	8302.2	25574.6	5187.1	14489.4	13859.6
λ_s	Min	36.6	50.6	71.5	48.1	89.1
	Median	386.7	566.1	436.2	607.8	1101.1
	Mean	405.7	607.4	479.1	626.6	1271.5
	Max	3969.7	5315.4	2449.5	4327.1	9429.4
λ_b	Min	37.4	41.7	52.8	47.6	71.9
	Median	388.2	610.3	418.4	645.8	1124.4
	Mean	411.8	651.6	456.4	665.3	1275.3
	Max	2335.3	5580.8	2730.8	3818.0	8406.7
PIN	Min	0.024	0.023	0.021	0.022	0.022
	Median	0.122	0.124	0.122	0.123	0.120
	Mean	0.123	0.124	0.121	0.123	0.121
	Max	0.316	0.295	0.301	0.342	0.326

Table 2: Summary of maximum likelihood estimates over the 3020 day sample period

Parameter	Statistic	BA	DIS	IBM	KO	XOM
α	Min	0.071	0.071	0.071	0.071	0.071
	Median	0.290	0.321	0.286	0.321	0.321
	Mean	0.309	0.322	0.306	0.321	0.321
	Max	0.929	0.929	0.928	0.929	0.837
δ	Min	0.044	0.043	0.037	0.039	0.041
	Median	0.349	0.339	0.334	0.373	0.250
	Mean	0.426	0.420	0.415	0.428	0.367
	Max	0.963	0.963	0.957	0.963	0.950
μ	Min	28.5	32.2	37.8	27.9	51.1
	Median	476.2	670.9	527.6	723.2	1245.6
	Mean	548.6	792.8	615.5	840.3	1456.3
	Max	8301.8	23529.0	5151.9	12332.3	13062.5
λ_s	Min	4.9	13.9	72.5	28.0	89.9
	Median	377.9	556.4	434.3	595.0	1084.9
	Mean	401.2	595.9	476.3	613.2	1253.1
	Max	3969.8	5315.5	2449.7	4327.0	9429.4
λ_b	Min	38.0	42.6	6.8	47.9	72.7
	Median	391.6	612.3	415.8	648.8	1125.2
	Mean	413.7	657.8	456.6	671.6	1284.7
	Max	2335.4	6126.8	2730.8	3306.1	8251.9
PIN	Min	0.055	0.047	0.039	0.052	0.039
	Median	0.138	0.135	0.134	0.137	0.127
	Mean	0.143	0.139	0.137	0.140	0.132
	Max	0.437	0.433	0.508	0.494	0.312

Table 3: Summary of Bayesian estimates over the 3020 day sample period

In our sample, the MLE approach of Yan and Zhang (2012) with the Lin and Ke (2011) factorization is successful in producing a PIN estimate for each day. However, the proportion of days on which δ is estimated as 0 or 1 (corner solutions) varies between 34% for KO and 50% for XOM. This is despite the fact that the size of inputs (numbers of buy and sell trades over 15−minute intervals) that are used for daily estimation is much smaller than those used over the typical two-three month period (using daily counts). In contrast, the Bayesian results show no trading day with corner solutions for δ, as expected.

We can review the results plotted over time (Figure 4), and also as a histogram (Figures 5 and 6). The histograms are plotted on the same scale for comparison. In each of these

figures, the data we are plotting are the daily estimates of PIN for each of the stocks. It is important to note that the daily values are estimated independent of each other. Yet, plotted as a time series, the contrast between the stability and smoothness of the two sets of estimates is clearly noticeable.

[Insert Figures 4, 5 and 6 near here]

Finally, we can quantify the daily variation in PIN estimates by calculating the size of the daily changes over the entire sample. We calculate the absolute value of the daily changes in PIN estimates for each stock, and then report the median, mean and variance of these changes in Table 4. The table confirms the relative stability of the Bayesian PIN estimates. The average (median or mean) size of the daily change in MLE PIN estimates is between 25% and 48% larger than that of the Bayesian PIN estimates. The variance of these changes is between 37% and 68% larger for the MLE estimates relative to the Bayesian estimates.

Statistic	Method	BA	DIS	IBM	KO	XOM
Median	MLE	0.04225	0.04143	0.03912	0.04253	0.04386
	BayesPin	0.03292	0.03073	0.03028	0.02880	0.03098
Mean	MLE	0.05101	0.04906	0.04776	0.0505	0.05180
	BayesPin	0.04089	0.03738	0.03659	0.03643	0.03765
Variance	MLE	0.00155	0.00143	0.00138	0.00150	0.00159
	BayesPin	0.00113	0.00095	0.00092	0.00095	0.00095

Table 4: Comparison of daily absolute changes in PIN estimates

4.2.1. Patterns of difference between MLE and PIN estimates

In order to understand the differences between the daily estimates from the MLE and Bayes algorithms, we plot the histogram of these differences. One potential source of difference identified by the literature is the occurrence of corner solutions in parameter estimates. To evaluate this source, we split the observations into two groups - one for which the MLE estimate of δ is either 0 or 1, the other for which it is not on the boundary. Looking at the

histograms for the five assets in Figure 7, we can see that there is a clear difference between the two sets of observations. On days when there is no corner solution, the Bayesian and ML estimates are more likely to agree, with their differences concentrated on zero and spread evenly around that value. On the other hand, the ML estimates are much more dispersed around the Bayesian estimates, and biased relative to these estimates, on days which δ is either 0 or 1. We have used a criterion based on a documented example of a numerical problem with the ML estimates, and can see clearly that this specific numerical problem leads to a relatively higher proportion of extreme differences between the two estimates. We argue this is supportive evidence for the relative reliability of the Bayesian estimates.

4.3. Quarterly PIN estimation results

In order to estimate a series of quarterly PIN values, we use aggregated buy and sell numbers over a day for all trading days in each calendar quarter. While large scale studies involving PIN use annual estimates (see, *e.g.*, Easley et al., 2002), other applications use frequencies as low as quarterly (see, *e.g.*, Christoffersen et al., 2018). The first PIN estimate is for January - March 2004, which uses all the trading days up to and including March 31, 2004. This procedure leads to 48 quarterly PIN estimates. Tables 5 and 6 provide summaries of the quarterly PIN estimates computed using the MLE and BayesPin approaches respectively.

The proportion of the quarterly MLE PIN estimates in which the estimate of δ is a corner solution (either 0 or 1) varies between 29% for BA and 48% for XOM. These proportions are similar to the daily case, although the size of the aggregate numbers of buys and sells is much larger over the course of a day, potentially leading to more computational issues. This also suggests that the size of the numbers may not necessarily be the only cause of numerical issues with the ML estimator. Given that there are 48 quarterly observations, potentially having half of the PIN estimates based on a corner solution is clearly a matter for concern.

As in the daily case, we plot the two series together over time in Figure 8, for each of the assets. In the figure, we also plot the 95% credible interval of PIN using the Bayesian

Parameter	Statistic	BA	DIS	IBM	KO	XOM
α	Min	0.016	0.016	0.033	0.115	0.109
	Median	0.315	0.276	0.254	0.347	0.432
	Mean	0.295	0.277	0.257	0.362	0.416
	Max	0.500	0.625	0.519	0.813	0.828
δ	Min	0.000	0.000	0.000	0.000	0.000
	Median	0.095	0.043	0.179	0.046	0.000
	Mean	0.316	0.224	0.324	0.149	0.150
	Max	1.000	1.000	1.000	1.000	1.000
μ	Min	611.8	712.6	867.6	490.5	798.9
	Median	8023.8	11705.6	10185.6	11531.7	16313.5
	Mean	10453.9	16762.6	10046.0	11298.8	19901.4
	Max	97221.6	87730.8	39088.0	29440.9	82423.1
λ_s	Min	1819.4	2246.7	3261.7	2187.3	3709.0
	Median	12046.6	18124.6	13963.9	20082.2	30902.5
	Mean	11077.5	16986.1	13041.6	17660.6	34933.1
	Max	25289.5	42301.8	33056.2	41275.3	130761.0
λ_b	Min	1972.5	2321.4	3046.1	2271.4	3478.4
	Median	12312.6	19019.3	12750.1	19278.7	32371.2
	Mean	11233.7	17325.4	12743.3	16986.2	32531.0
	Max	23365.2	39863.6	30985.3	35381.7	87822.2
PIN	Min	0.010	0.011	0.021	0.047	0.055
	Median	0.096	0.088	0.085	0.103	0.106
	Mean	0.099	0.094	0.085	0.115	0.120
	Max	0.170	0.195	0.148	0.305	0.311

Table 5: Summary of quarterly maximum likelihood estimates over sample period

approach. Although the two series agree on the direction of changes in many cases, we can see that in the case of the latter half of the sample for BA and DIS, there are some extreme opposing spikes in the MLE PIN. For completeness, we again report absolute changes and variance over time of the two sets of estimates for each stock in Table 7, although at the quarterly level, it is not our goal to emphasize relative stability to the same extent. The average (median or mean) size of the daily change in MLE PIN estimates is between 31% and 63% larger than that of the Bayesian PIN estimates. The variance of these changes is between 74% and 139% larger for the MLE estimates relative to the Bayesian estimates.

More so, Figure 8 suggests that there are certain quarters when the MLE estimates are

Parameter	Statistic	BA	DIS	IBM	KO	XOM
α	Min	0.045	0.079	0.076	0.141	0.121
	Median	0.328	0.333	0.267	0.354	0.428
	Mean	0.312	0.323	0.274	0.369	0.402
	Max	0.498	0.562	0.522	0.636	0.727
δ	Min	0.031	0.030	0.033	0.027	0.020
	Median	0.194	0.087	0.207	0.136	0.048
	Mean	0.321	0.226	0.317	0.270	0.199
	Max	0.967	0.913	0.965	0.977	0.972
μ	Min	611.8	713.9	874.2	490.5	798.9
	Median	7765.2	10932.9	9577.9	11101.9	15953.7
	Mean	9200.9	13252.1	9577.4	10875.2	19199.0
	Max	60115.3	87731.0	39084.2	28440.0	67034.5
λ_s	Min	1819.4	2246.5	3261.6	2183.3	3708.9
	Median	11917.4	18210.1	13917.9	19592.8	32028.9
	Mean	11090.3	16960.5	13233.0	17055.6	34243.4
	Max	25289.7	42301.5	33056.2	31361.1	101370.2
λ_b	Min	1970.3	2322.5	3043.8	2271.6	3479.7
	Median	12300.2	18117.2	13277.1	19231.3	31644.8
	Mean	11180.5	16997.8	12506.2	17670.4	34151.3
	Max	23365.3	39863.6	27580.2	46149.5	137345.2
PIN	Min	0.055	0.041	0.036	0.042	0.044
	Median	0.083	0.088	0.075	0.080	0.079
	Mean	0.089	0.092	0.075	0.092	0.090
	Max	0.136	0.192	0.114	0.197	0.191

Table 6: Summary of quarterly Bayesian estimates over sample period

not only outside the credible interval, they also have sharp drops to values near zero or sharp increases. It is, however, reassuring that the two methods do tend to move similarly except in the extreme cases identified above.

[Insert Figure 8 near here]

Statistic	Method	BA	DIS	IBM	KO	XOM
Median	MLE	0.02772	0.03753	0.02612	0.04202	0.03665
	BayesPin	0.01911	0.02298	0.01694	0.02958	0.02793
Mean	MLE	0.03288	0.04474	0.03013	0.05237	0.05207
	BayesPin	0.02249	0.02958	0.01968	0.03941	0.03651
Variance	MLE	0.00060	0.00140	0.00038	0.00273	0.00262
	BayesPin	0.00025	0.00080	0.00022	0.00125	0.00123

Table 7: Comparison of absolute changes in quarterly PIN estimates

5. Simulation Study

In this section we carry out two simulation exercises to evaluate whether the Bayesian estimation performs comparably to the MLE algorithm. As we use the *InfoTrad* package (see Çelik and Tiniç, 2018) for comparison to the Bayesian algorithm, we follow the format of the exercises provided alongside the package.

5.1. Fixed PIN at varying trade intensities

For the first simulation, we generate $1,000$ sets each of 60 observations (corresponding to one quarter), at varying trade intensities with the following parameters: $\alpha = 0.5, \delta = 0.5, \mu = 0.2k, \lambda_b = \lambda_s = 0.4k$, where $k = \{100, 500, 1000, ..., 5000\}$. The values of these parameters are adopted from the exercises published by Çelik and Tiniç (2018) and Gan et al. (2015), in order to be as consistent as possible with the existing methodology. For each trading period, we draw a value from the binomial distribution with parameter α and δ respectively to determine whether it was a good news day, bad news day, or no news day. We then draw Poisson random values corresponding to the relevant intensity levels for each case, using the parameters μ, λ_b and λ_s. Thus we have a total of $11,000$ simulated sample sets. Scatter plots of the buy and sell trades generated from this simulation at 6 different intensities are shown in Figure 9. We can see, in comparison to the scatter plots in Figure 2,

the difference between the empirical and theoretical versions of the Buy-Sell pairs. In order to discriminate between different levels of α, we also repeat the above exercise with $\alpha = 0.25$ and $\alpha = 0.75$.

For each 60 day sample, we estimate PIN using the BayesPIN toolbox, as well as the *InfoTrad* package with various specifications. We specify the MLE estimation with the Gan et al. (2015) algorithm and the Yan and Zhang (2012) approach, each with both the Lin and Ke (2011) and the Easley et al. (2002) likelihood factorization (GWJ-LK, GWJ-EHO, YZ-LK, and YZ-EHO). We use all combinations of the specifications because, by design, we should expect the clustering based methods (GWJ) to work best based on the data we have generated.

The overall performance results are in Table 8. We can see that the BayesPIN algorithm performs favorably with the best of the MLE alternatives, viz. the clustering algorithm of Gan et al. (2015) combined with the Lin and Ke (2011) factorization (GWJ-LK). In general, this factorization works better in the MLE case than all others. As expected, the Yan and Zhang (2012) approach also performs well (YZ-LK). It is also interesting that there are far fewer corner solutions in the MLE results with the simulated data, suggesting that we should treat the performance from simulations with a small degree of caution. In the case of $\alpha = 0.5$, only the YZ-EHO algorithm generates 158 values of 0 or 1 for α or δ, out of $11,000$ cases. Similarly for $\alpha = 0.75$, it produces 191 such estimates. In the case of $\alpha = 0.25$, apart from the 150 corner solutions produced by the YZ-EHO algorithm, the GWJ-EHO, GWJ-LK, and YZ-LK respectively produce 6, 4, and 22 corner estimates.

In Figure 10, we show the histogram of PIN estimates for $\alpha = 0.5$ at varying intensities (each containing $1,000$ simulated samples as described above) for the Bayesian algorithm alongside the GWJ-LK estimates. The Bayesian algorithm performs well at all intensities.

	Method	Statistic	α	δ	μ	λ_s	λ_b	PIN
$\alpha = 0.25$								
	GWJ-LK	MSE	0.0040	0.0189	116.4	18.7	18.2	0.0002
		MAE	0.0476	0.1083	8.1	3.2	3.2	0.0104
	GWJ-EHO	MSE	0.0072	0.0247	49585.4	75135.8	54092.9	0.0018
		MAE	0.0639	0.1215	129.5	152.2	134.2	0.0254
	YZ-LK	MSE	0.0039	0.0189	116.4	18.6	18.3	0.0006
		MAE	0.0476	0.1083	8.1	3.2	3.2	0.0203
	YZ-EHO	MSE	0.4877	0.1951	143111.6	912902.1	1587633.3	0.0353
		MAE	0.6928	0.4381	264.6	813.8	978.7	0.1374
	BayesPIN	MSE	0.0036	0.0137	116.3	18.6	18.2	0.0002
		MAE	0.0467	0.0931	8.1	3.2	3.2	0.0100
$\alpha = 0.5$								
	GWJ-LK	MSE	0.0045	0.0089	62.6	20.8	20.4	0.0002
		MAE	0.0529	0.0751	5.9	3.4	3.4	0.0107
	GWJ-EHO	MSE	0.0075	0.0120	44640.9	32470.4	57796.5	0.0031
		MAE	0.0676	0.0850	121.2	121.9	136.8	0.0338
	YZ-LK	MSE	0.0045	0.0089	62.6	20.8	20.4	0.0013
		MAE	0.0529	0.0751	5.9	3.4	3.4	0.0319
	YZ-EHO	MSE	0.2045	0.1899	131656.3	873448.6	1410405.9	0.0213
		MAE	0.4467	0.4308	264.2	792.6	856.9	0.1067
	BayesPIN	MSE	0.0042	0.0077	62.5	20.8	20.4	0.0002
		MAE	0.0510	0.0701	5.9	3.4	3.4	0.0103
$\alpha = 0.75$								
	GWJ-LK	MSE	0.0660	0.0056	47.9	22.8	23	0.0023
		MAE	0.2501	0.0598	5.1	3.6	3.6	0.0467
	GWJ-EHO	MSE	0.0789	0.0090	36071.2	37459.9	63907.4	0.0071
		MAE	0.2711	0.0735	113.1	132.5	144.5	0.0610
	YZ-LK	MSE	0.0660	0.0056	47.9	22.8	23	0.0080
		MAE	0.2501	0.0598	5.1	3.6	3.6	0.0887
	YZ-EHO	MSE	0.2082	0.1896	113532.4	830222.1	1035560.4	0.0204
		MAE	0.4538	0.4296	249.5	768.1	702.8	0.1066
	BayesPIN	MSE	0.0615	0.0051	47.9	22.8	23.0	0.0021
		MAE	0.2415	0.0571	5.1	3.6	3.6	0.0450

Table 8: Comparison of parameter estimation methods for simulated datasets with varying trading intensities, at three levels of α and fixed $\delta = 0.5$

5.2. Randomly generated parameters

The second exercise is to simulate $5,000$ random combinations of the parameters using the following rules: First generate 3 sets of uniform random numbers for each of the $5,000$

data sets we wish to generate. We use the first random number to represent α, the second to represent δ, and the third to represent μ, whereby $\lambda_b = \lambda_s = 0.5(1-\mu)$. The last three parameters are multiplied by $2,500$ to obtain the respective intensities. Scatter plots of 6 randomly chosen samples from the $5,000$ generated are provided in Figure 11 to demonstrate the range of possible scenarios.

Method		Statistic	α	δ	μ	λ_s	λ_b	PIN
GWJ	LK	MSE	0.0203	0.0232	34881.6	948.6	163.6	0.0012
		MAE	0.0644	0.0896	31.4	4.3	3.8	0.0182
	EHO	MSE	0.0633	0.0789	1165664	124492.9	202829.8	0.1149
		MAE	0.1655	0.1943	807.2	258	330.4	0.2312
YZ	LK	MSE	0.0178	0.0231	35696.1	170	116.4	0.0583
		MAE	0.0607	0.0896	30.2	3.5	3.5	0.1598
	EHO	MSE	0.2767	0.1775	1254183.1	268256.4	792743.2	0.0941
		MAE	0.4400	0.3610	910.9	400.3	668.1	0.2309
BayesPIN		MSE	0.0101	0.0134	34312.7	284.8	128.1	0.0010
		MAE	0.0535	0.0770	35.5	3.7	3.5	0.0176

Table 9: Performance of alternative estimation algorithms on simulated data

Once again, we see that the Bayesian algorithm is comparable to the MLE estimators (see Table 9 and Figure 12). Note that in these results we have excluded three cases because the YZ-EHO and YZ-LK algorithms did not produce an estimate. In addition, the two LK factorization algorithms still produced values of either 0 or 1 for α or δ. Specifically, GWJ-LK produced 35 corner solutions for α and 209 for δ, while the YZ-LK produced respectively 101 and 538 corner solutions.

6. Conclusion

We have shown a Bayesian method that can provide estimates of PIN using small data sets at higher frequency, as well as over longer periods for stocks with very large numbers of trades. This method avoids the non-convergence and other computational problems of optimization functions underlying MLE routines. We know from previous literature that

two of the challenging parameters to estimate using MLE are α (the probability of news arrival) and δ (the probability that the news is bad). We have also found that the MLE algorithm gives us a boundary value of either zero or one in a considerable number of cases, particularly for δ. The Bayesian methodology, on the other hand, does not suffer from this corner solution problem. The Bayesian approach is also not sensitive to the initial values in the way that is known to be the case for the MLE approach.

One consequence of the MLE estimation problem is that independently estimated values of PIN for each day could swing back and forth over time on account of numerical issues. In the empirical illustrations, we demonstrated that the Bayesian estimation works well even at a daily frequency using intraday data sampled over only 26 trading periods of 15−minutes each. The ability to estimate PIN in this manner offers new opportunities to apply the measure in studies involving time-varying information asymmetry risk.

References

Aktas, N., De Bodt, E., Declerck, F., Van Oppens, H., 2007. The PIN anomaly around M&A announcements. Journal of Financial Markets 10, 169–191.

Boehmer, E., Grammig, J., Theissen, E., 2007. Estimating the probability of informed trading: Does trade misclassification matter? Journal of Financial Markets 10, 26–47.

Brennan, M. J., Huh, S.-W., Subrahmanyam, A., 2018. High-frequency measures of informed trading and corporate announcements. The Review of Financial Studies 31, 2326–2376.

Çelik, D., Tiniç, M., 2018. Infotrad: An r package for estimating the probability of informed trading. R Journal 10, 31–42.

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

Christoffersen, P., Goyenko, R., Jacobs, K., Karoui, M., 2018. Illiquidity premia in the equity options market. The Review of Financial Studies 31, 811–851.

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? The 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. The Journal of Finance 51, 1405–1436.

Easley, D., O'Hara, M., 1987. Price, trade size, and information in securities markets. Journal of Financial Economics 19, 69–90.

Easley, D., O'Hara, M., 1992. Adverse selection and large trade volume: The implications for market efficiency. Journal of Financial and Quantitative Analysis 27, 185–208.

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

Gan, Q., Wei, W. C., Johnstone, D., 2015. A faster estimation method for the probability of informed trading using hierarchical agglomerative clustering. Quantitative Finance 15, 1805–1821.

Gan, Q., Wei, W. C., Johnstone, D., 2017. Does the probability of informed trading model fit empirical data? Financial Review 52, 5–35.

Glosten, L. R., Milgrom, P. R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 14, 71–100.

Holden, C. W., Jacobsen, S., 2014. Liquidity measurement problems in fast, competitive markets: expensive and cheap solutions. The Journal of Finance 69, 1747–1785.

Jackson, D., 2013. Estimating PIN for firms with high levels of trading. Journal of Empirical Finance 24, 116–120.

Korajczyk, R. A., Sadka, R., 2008. Pricing the commonality across alternative measures of liquidity. Journal of Financial Economics 87, 45–72.

Kyle, A., 1985. Continuous auctions and insider trading. Econometrica: Journal of the Econometric Society pp. 1315–1335.

Lai, S., Ng, L., Zhang, B., 2014. Does PIN affect equity prices around the world? Journal of Financial Economics 114, 178–195.

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

Lei, Q., Wu, G., 2005. Time-varying informed and uninformed trading activities. Journal of Financial Markets 8, 153–181.

Lin, W. H.-W., Ke, W.-C., 2011. A computing bias in estimating the probability of informed trading. Journal of Financial Markets 14, 625–640.

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

Van Dyk, D. A., Meng, X.-L., 2001. The art of data augmentation. Journal of Computational and Graphical Statistics 10, 1–50.

Vega, C., 2006. Stock price reaction to public and private information. Journal of Financial Economics 82, 103 – 133.

Venter, J. H., De Jongh, D. C., 2006. Extending the EKOP model to estimate the probability of informed trading. Studies in Economics and Econometrics 30, 25–39.

Yan, Y., Zhang, S., 2012. An improved estimation method and empirical properties of the probability of informed trading. Journal of Banking & Finance 36, 454–467.

Appendix A. The BayesPin Toolbox

In what follows we provide details of the **BayesPin** toolbox written in **Matlab** that accompanies this paper. The toolbox calculates PIN based on the Easley et al. (2002), and Easley et al. (1996) models. The command for invoking the toolbox is

```
BayesPin(trades,model,sweeps,burnin,confidence).
```

The inputs of the toolbox are described as follows

- **trades**: a dataframe holding the aggregate buy and sell trades.
- **model**: Either of the Easley et al. (1996), Easley et al. (2002) models (i.e `EKOP96,EHO2002`).
- **sweeps (optional)**: the number of iterations for the Gibbs Sampler (the default is set at $25,000$). This has to be large enough to ensure convergence of the Markov chain. In our example, we used $10,000$.
- **burnin (optional)**: This is the number of initial iterations for which the parameter draws should be discarded. This is to ensure that we keep the draws at the point where the MCMC has converged to the parameter space in which the parameter estimate is likely to fall. This figure `must always` be less than the sweeps (the default is set at $5,000$).
- **confidence (optional)**: A number indicating the level of tolerance for computation of credible interval (the default is 5 for a 95% credible interval)

The output of the toolbox will be a list of the following

- The posterior estimates of the model parameters and the PIN.
- The standard deviations of the posterior draws of parameters and PIN.
- The lower credible limit of the posterior distribution for each parameter and PIN.
- The upper credible limit of the posterior distribution for each parameter and PIN.
- A Geweke statistic for each parameter and PIN.
- A p-value of the Geweke statistic for each parameter and PIN.

- The stored posterior distribution of each parameter and PIN as a matrix.

We provide below an illustrative example of daily parameter estimates and PIN for the Easley et al. (2002) model.

A.1. Matlab Implementation

In what follows it is assumed that the user of the toolbox have saved the folders of the toolbox into a personal folder with sub-folder *Results*.

```
%#########################################################################
%   EXAMPLE  : Calculation of Cross-Sectional PIN
%#########################################################################
clc
dir=[pwd,'\',mfilename]; %Working directory
cd (dir)    %Change to working directory

%import aggregated buy and sell trades
trades     = importdata(‘sampleData.txt’);
model      = ‘EHO2002’; % PIN Model of interest (i.e EKOP96, EHO2002)
sweeps     = 10000;      %Specify number of iterations
burnin     = 1000;       %This has to be smaller than sweeps
confidence = 5;          %Confidence level for credible interval
[est, stored_parameter_draws]= BayesPin(trades, model, sweeps, burnin, confidence);

est;       %View the estimates
%Save summary of posterior estimates over estimation period
filename = strcat(dir, '\Results\', model, '_CrossSection_PIN', '.csv');
writetable(est, filename, 'Delimiter', ',', 'QuoteStrings', true, 'WriteVariableNames',
true, 'WriteRowNames', true)

% Inspect the posterior distributions and trace plots of parameters and PIN
if string(model)=='EKOP96'    labs =  {'\alpha', '\delta', '\mu', '\epsilon', 'PIN'};
else    labs = {'\alpha', '\delta', '\mu', '\lambda_{s}', '\lambda_{b}','PIN'};
end
```

```
N = size(stored_parameter_draws, 2)
nbins=20;

%save plot as jpeg
hf=figure('Visible','on');
for i=1:N
subplot(2,N,i)
hist(stored_parameter_draws(:,i),nbins);axis tight
title(labs{i})
subplot(2,N,i+N)
plot(stored_parameter_draws(:,i));axis tight
title(labs{i})
end
saveas(hf,strcat(dir,'\Results\',model,'_PosteriorDist_Tracer.jpg'))
```

Appendix B. Derivation of joint density and full conditionals

In this appendix, we provide the derivations of the posterior density and full conditionals for the parameters.

B.1. Joint density

First, we write down the joint densities of buy and sell orders conditional on the realization of D_t.

Bad News Event ($D_t = 1$)

The model assumes that the number of sell trades from informed traders and the total number of sell trades are both generated from Poisson distributions. It can easily be shown that conditioning on the total number of sell trades, the number of sell trades by informed traders follows a binomial distribution with S_t trials and success probability $\mu/\mu+\lambda_s$. Sell trades initiated by uninformed traders are then calculated as $S_t^u = S_t - S_t^i$. The trade arrival distributions in the event of bad news are therefore given as

$$S_t|D_t = 1 \sim Pn\left(\mu + \lambda_s\right) \qquad B_t|D_t = 1 \sim Pn\left(\lambda_b\right),$$
$$S_t^i|S_t, D_t = 1 \sim Bin\left(S_t, \tfrac{\mu}{\mu+\lambda_s}\right).$$

The probability of the numbers of different types of buy or sell trades in the event of bad

news is

$$
\begin{aligned}
f_1\left(B_t, S, S_t^i, \Theta\right) &= P\left(B_t, S_t, S_t^i | D_t = 1, \Theta\right) \\
&= P\left(B_t | D_t = 1, \Theta\right) P\left(S_t^i | S_t, D_t = 1, \Theta\right) P\left(S_t | D_t = 1, \Theta\right) \\
&= \binom{S_t}{S_t^i} \frac{e^{-(\mu+\lambda_b+\lambda_S)}}{B_t! S_t!} {\lambda_b}^{B_t} {\lambda_S}^{S_t} \left(\frac{\mu}{\lambda_S}\right)^{S_t^i} \\
&= \frac{e^{-\lambda_b}\left(\lambda_b\right)^{B_t}}{B_t!} \binom{S_t}{S_t^i} \left(\frac{\mu}{\mu+\lambda_S}\right)^{S_t^i} \left(\frac{\lambda_S}{\mu+\lambda_S}\right)^{S_t - S_t^i} \frac{e^{-\mu+\lambda_{S_t}}(\mu+\lambda_S)^{S_t}}{S_t!} \\
&= \binom{S_t}{S_t^i} \frac{e^{-(\mu+\lambda_b+\lambda_S)}}{B_t! S_t!} {\lambda_b}^{B_t} {\lambda_S}^{S_t - S_t^i} \mu^{S_t^i} \\
&= \frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!} \frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!} \frac{e^{-\lambda_S}\lambda_S^{S_t - S_t^i}}{(S_t - S_t^i)!}.
\end{aligned} \tag{6}
$$

As expected, Equation 6 is a product of Poisson processes for the buy trades and sell trades by uninformed traders, and the sell trades by informed traders.

Good News Event ($D_t = 2$)

Using similar arguments to those above, the distributions of the numbers of different types of trades in the event of good news can be written as follows

$$
\begin{aligned}
&B_t | D_t = 2 \sim Pn\left(\mu + \lambda_b\right), && S_t | D_t = 2 \sim Pn\left(\lambda_s\right), \\
&B_t^i | B_t, D_t = 2 \sim Bin\left(B_t, \tfrac{\mu}{\mu+\lambda_b}\right).
\end{aligned}
$$

The probability of the numbers of different types of buyer or seller initiated trades is

$$
\begin{aligned}
f_2\left(B_t, S_t, \Theta\right) &= P\left(B_t, S_t | D_t = 2, \Theta\right) \\
&= P\left(S_t | D_t = 2, \Theta\right) P\left(B_t^i | B_t, D_t = 2, \Theta\right) P\left(B_t | D_t = 2, \Theta\right) \\
&= \frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!} \frac{e^{-\lambda_s}\lambda_s^{S_t}}{S!} \frac{e^{-\lambda_b}\lambda_b^{B_t - B_t^i}}{(B_t - B_t^i)!}.
\end{aligned} \tag{7}
$$

No News Event ($D_t = 3$)

With the assumption that there is no informed trader activity during a "no news" period, all trades are attributable to uninformed traders. Hence we have $B_t|D_t = 3 \sim Pn(\lambda_b)$ and $S_t|D_t = 3 \sim Pn(\lambda_s)$ as the distributions of the number of buys and sells respectively. The probability of the number of buyer and seller initiated trades is

$$\begin{aligned} f_3(B_t, S_t, \Theta) =& P(B_t, S_t|D_t = 3, \Theta) \\ =& P(S_t|D_t = 3, \Theta) P(B_t|\Theta) P(B_t|D_t = 3, \Theta) \\ =& \frac{e^{-\lambda_s}(\lambda_s)^{S_t}}{S_t!} \frac{e^{-\lambda_b}(\lambda_b)^{B_t}}{B_t!}. \end{aligned} \tag{8}$$

Putting Equations 6, 7 and 8 together, we obtain the joint probability function

$$\begin{aligned} P(B_t, S_t|D_t, \Theta) =& \left[f_1(B_t, S_t, \Theta) \right]^{d_{t,1}} \left[f_2(B_t, S_t, \Theta) \right]^{d_{t,2}} \left[f_3(B_t, S_t, \Theta) \right]^{d_{t,3}} \\ =& \left[\frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!} \frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!} \frac{e^{-\lambda_s}\lambda_s^{S_t - S_t^i}}{(S_t - S_t^i)!} \right]^{d_{t,1}} \left[\frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!} \frac{e^{-\lambda_s}\lambda_s^{S_t}}{S_t!} \frac{e^{-\lambda_b}\lambda_b^{B_t - B_t^i}}{(B_t - B_t^i)!} \right]^{d_{t,2}} \\ & \times \left[\frac{e^{-\lambda_s}(\lambda_s)^{S_t}}{S_t!} \frac{e^{-\lambda_b}(\lambda_b)^{B_t}}{B_t!} \right]^{d_{t,3}}, \end{aligned} \tag{9}$$

B.2. Derivation of Posterior Distributions

Posterior Density

In Bayesian inference, the posterior distribution is proportional to the product of the prior distribution and the likelihood function of the parameters. Using the prior distributions provided in Section 3.3.1 and the likelihood function, we obtain the posterior density of buys and sells:

$P\left(D_t, \Theta | B_t, S_t\right)$ is proportional to

$$
\begin{aligned}
& P\left(\Theta\right) \prod_{t=1}^{T} \left[P\left(B_t, S_t | D_t, \Theta\right) P\left(D_t | \Theta\right) \right] \\
& = P\left(\Theta\right) \prod_{t=1}^{T} \left[\left[\alpha\delta \frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!} \frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!} \frac{e^{-\lambda_s}\lambda_s^{S_t - S_t^i}}{(S_t - S_t^i)!} \right]^{d_1} \left[\alpha(1-\delta) \frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!} \frac{e^{-\lambda_s}\lambda_s^{S_t}}{S_t!} \frac{e^{-\lambda_b}\lambda_b^{B_t - B_t^i}}{(B_t - B_t^i)!} \right]^{d_2} \right. \\
& \left. \times \left[(1-\alpha) \frac{e^{-\lambda_s}\left(\lambda_s\right)^{S_t}}{S_t!} \frac{e^{-\lambda_b}\left(\lambda_b\right)^{B_t}}{B_t!} \right]^{d_3} \right] \\
& = P\left(\Theta\right) \left[(\alpha\delta)^{T_1} (\alpha(1-\delta))^{T_2} (1-\alpha)^{T_3} \right] \\
& \times \prod_{t=1}^{T} \left[\frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!} \frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!} \frac{e^{-\lambda_s}\lambda_s^{S_t - S_t^i}}{(S_t - S_t^i)!} \right]^{d_1} \left[\frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!} \frac{e^{-\lambda_s}\lambda_s^{S_t}}{S_t!} \frac{e^{-\lambda_b}\lambda_b^{B_t - B_t^i}}{(B_t - B_t^i)!} \right]^{d_2} \left[\frac{e^{-\lambda_s}\left(\lambda_s\right)^{S_t}}{S_t!} \frac{e^{-\lambda_b}\left(\lambda_b\right)^{B_t}}{B_t!} \right]^{d_3} \\
& = \eta^{2a+b} \mu^{b-1} e^{-\eta\mu} \lambda_s^{a-1} e^{-\eta\lambda_s} \lambda_b^{a-1} e^{-\eta\lambda_b} \eta^{0.001-1} e^{-0.001\eta} \alpha^1 (1-\alpha)^1 \delta^1 (1-\delta)^1 \left[(\alpha\delta)^{T_1} (\alpha(1-\delta))^{T_2} (1-\alpha)^{T_3} \right] \\
& \times \prod_{t=1}^{T} \left[\frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!} \frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!} \frac{e^{-\lambda_s}\lambda_s^{S_t - S_t^i}}{(S_t - S_t^i)!} \right]^{d_1} \left[\frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!} \frac{e^{-\lambda_s}\lambda_s^{S_t}}{S_t!} \frac{e^{-\lambda_b}\lambda_b^{B_t - B_t^i}}{(B_t - B_t^i)!} \right]^{d_2} \left[\frac{e^{-\lambda_s}\left(\lambda_s\right)^{S_t}}{S_t!} \frac{e^{-\lambda_b}\left(\lambda_b\right)^{B_t}}{B_t!} \right]^{d_3}
\end{aligned}
\tag{10}
$$

Full Conditional Distributions

In order to use the Gibbs sampler we need the full conditionals of the parameters of interest. In order to use the Gibbs sampler we need the full conditionals of the parameters of interest, $\Theta = (\alpha, \delta, \mu, \lambda_s, \lambda_s, \eta)$. Using Equation 10 we derive below the full conditional distributions of the parameters.

Full Conditional for news impact parameter η

Similarly, the full conditional distribution for η is derived as follows :

$$
\begin{aligned}
P(\eta|.) &\propto \eta^{2a+b} \mu^{b-1} e^{-\eta\mu} \lambda_s^{a-1} e^{-\eta\lambda_s} \lambda_b^{a-1} e^{-\eta\lambda_b} \eta^{0.001-1} e^{-0.001\eta} \\
&\propto \eta^{0.001+2a+b} e^{-\eta(0.001+\mu+\lambda_s+\lambda_b)}
\end{aligned}
\tag{11}
$$

The expression in equation 11 is the kernel of a gamma distribution therefore the full

conditional distribution of $\eta \sim \mathcal{G}a\left(0.001 + 2a + b, \quad 0.001 + \mu + \lambda_s + \lambda_b\right)$.

Full Conditional for informed trader arrivals parameter $\boldsymbol{\mu}$

$$\begin{aligned} P(\mu|.) \quad &\propto \mu^{b-1} e^{-\eta\mu} \prod_{t=1}^{T} \left[\frac{e^{-\mu}\mu^{S_t^i}}{S_t^i!}\right]^{d_1} \left[\frac{e^{-\mu}\mu^{B_t^i}}{B_t^i!}\right]^{d_2} \\ &\propto \frac{e^{-\mu(T_1+T_2)}\mu^{b-1}e^{-\eta\mu}\mu^{\sum_{t=1}^{T}\left(S_t^{id_1}+B_t^{id_2}\right)}}{\prod_{t=1}^{T}\left[S_t^i!\right]^{d_1}\left[B_t^i!\right]^{d_2}} \\ &\propto e^{-\mu(T_1+T_2)}\mu^{b-1}e^{-\eta\mu}\mu^{\sum_{t=1}^{T}\left(S_t^{id_1}+B_t^{id_2}\right)} \\ &\propto e^{-\mu(T_1+T_2+\eta)}\mu^{b+\sum_{t=1}^{T}\left(S_t^{id_1}+B_t^{id_2}\right)-1} \end{aligned} \tag{12}$$

This is the kernel of a gamma distribution hence the full conditional distribution of μ can be written as $\mu \sim \mathcal{G}a\left(b + \sum_{t=1}^{T}\left(S_{jt}^{id_1} + B_{jt}^{id_2}\right),\ \eta + T_1 + T_2\right)$.

Full Conditional for uninformed buy trader arrivals parameter $\boldsymbol{\lambda_b}$

$$\begin{aligned} P(\lambda_b|.) \quad &\propto e^{-\eta\lambda_b}\lambda_b^{a-1}\prod_{t=1}^{T}\left[\frac{e^{-\lambda_b}\lambda_b^{B_t}}{B_t!}\right]^{d_1}\left[\frac{e^{-\lambda_b}\lambda_b^{B_t-B_t^i}}{(B_t-B_t^i)!}\right]^{d_2}\left[\frac{e^{-\lambda_b}(\lambda_b)^{B_t}}{B_t!}\right]^{d_3} \\ &= \frac{e^{-\lambda_b\eta}\lambda_b^{a-1}e^{-(T_1+T_2+T_3)\lambda_b}\lambda_b^{\sum_{t=1}^{T}\left(B_t^{d_1}+B_t^{d_2}-B_t^{id_2}+B_t^{d_3}\right)}}{\prod_{t=1}^{T}\left[B_t!\right]^{d_1}\left[(B_t-B_t^i)!\right]^{d_2}\left[B!\right]^{d_3}} \\ &\propto e^{-(T_1+T_2+T_3+\eta)\lambda_b}\lambda_b^{\sum_{t=1}^{T}\left(B_t^{d_1}+B_t^{d_2}-B_t^{id_2}+B_t^{d_3}\right)+a-1} \end{aligned} \tag{13}$$

As before this full conditional distribution has the kernel of a gamma distribution therefore $\lambda_b \sim \mathcal{G}a\left(a + \sum_{t=1}^{T}\left(B_t^{d_1} + B_t^{d_2} - B_t^{id_2} + B_t^{d_3}\right),\quad \eta + T_1 + T_2 + T_3\right)$.

Full Conditional for uninformed sell trader arrivals parameter $\boldsymbol{\lambda_s}$

$$\begin{aligned} P(\lambda_s|.) \quad &\propto \lambda_s^{a-1} e^{-\eta\lambda_s} \prod_{t=1}^{T} \left[\frac{e^{-\lambda_s}\lambda_s^{S_t - S_t^i}}{(S_t - S_t^i)!}\right]^{d_1} \left[\frac{e^{-\lambda_s}\lambda_s^{S_t}}{S_t!}\right]^{d_2} \left[\frac{e^{-\lambda_s}(\lambda_s)^{S_t}}{S_t!}\right]^{d_3} \\ &= \frac{\lambda_s^{a-1} e^{-\eta\lambda_s} e^{-(T_1+T_2+T_3)\lambda_s} \lambda_s^{\sum_{t=1}^{T}\left(S_t^{d_1} - S_t^{id_1} + S_t^{d_2} + S_t^{d_3}\right)}}{\prod_{t=1}^{T}\left[(S_t - S_t^i)!\right]^{d_1}\left[S_t!\right]^{d_2}\left[S_t!\right]^{d_3}} \\ &\propto e^{-\lambda_s(T_1+T_2+T_3+\eta)} \lambda_s^{\sum_{t=1}^{T}\left(S_t^{d_1} - S_t^{id_1} + S_t^{d_2} + S_t^{d_3}\right)+a-1} \end{aligned} \tag{14}$$

The expression in equation 14 is the kernel of a gamma distribution therefore the full conditional distribution of $\lambda_s \sim \mathcal{G}a\left(a + \sum_{t=1}^{T}\left(S_t^{d_1} - S_t^{id_1} + S_t^{d_2} + S_t^{d_3}\right),\quad \eta + T_1 + T_2 + T_3\right)$.

Full Conditional for news arrival parameter $\boldsymbol{\alpha}$

The full conditional distribution for α, $P(\alpha|.)$, is derived as follows

$$\begin{aligned} P(\alpha|.) &\propto \alpha^{T_1}(1-\alpha)^{T_2}\alpha^1(1-\alpha)^1 \\ &\propto \alpha^{T_1+T_2+1}(1-\alpha)^{T_2+1}, \end{aligned} \tag{15}$$

which is the kernel of a beta distributions

$$\alpha \sim \mathcal{B}e\Big(1 + T_1 + T_2,\ T_3 + 1\Big)$$

Full Conditional for news impact parameter $\boldsymbol{\delta}$

Similarly, the full conditional distribution for δ is derived as follows :

$$\begin{aligned} P(\delta|.) &\propto \delta^{T_1}(1-\delta)^{T_2}\delta^1(1-\delta)^1 \\ &\propto \delta^{T_1+1}(1-\delta)^{T_2+1}. \end{aligned} \tag{16}$$

The expression in Equation 16 is the kernel of a beta distribution so

$$\delta \sim \mathcal{B}e\Big(1 + T_1,\ \ 1 + T_2\Big)$$

It can be observed that the estimates of α and δ are highly dependent on the correct classification of news event period.

Appendix C. Additional simulation exercise

To produce an alternative dataset that is not directly generated from the EHO-PIN model, we rely on an insight in Venter and De Jongh (2006), who introduced the Poisson Inverse Gaussian distribution in the EKOP-PIN model. To produce buy and sell trade pairs that are correlated but conditionally independent, we use a random unit inverse gaussian scaling factor on the trade arrival intensity parameters.

We generate $1,000$ sets each of 60 observations (corresponding to one quarter), at varying trade intensities with the following parameters: $\alpha = 0.5, \delta = 0.5, \mu = 0.2k, \lambda_b = \lambda_s = 0.4k$, where $k = \{100, 500, 1500, ..., 8500\}$. For each trading period, we draw a value from the binomial distribution with parameter α and δ respectively to determine whether it was a good news day, bad news day, or no news day. For each draw of Poisson trader arrivals in a given trading period in a particular sample path, we first scale all three intensity parameters $(\mu, \lambda_b, \lambda_s)$ by a random draw (each trading period) from an inverse gaussian distribution with expected value 1 and scale parameter 9. Note that we rescale the random inverse gaussian draws in order to ensure the mean of the scaling factor on each sample path is 1. We then draw Poisson random values corresponding to the relevant scaled intensity levels for each case. Thus we have a total of $10,000$ simulated sample sets. The data generated from this model still produces distinct clusters, but less so than those produced by the EHO-PIN model. In Figure 13, we present scatter plots of buy and sell pairs from 6 samples from the simulated dataset. We then estimate the PIN model on the simulated data from this exercise, using both MLE and BayesPIN. As in the main simulation, we repeat the exercise with two other values for $\alpha \in \{0.25, 0.75\}$.

We naturally expect the results to be biased, but we find that the Bayesian and the best ML estimators still discriminate between the three different levels of PIN (using t-tests and ranksum tests not reported here, but available from the authors). The Bayesian estimator again performs very well, providing additional simulation evidence of the features we have observed in the empirical estimation.

	Method	Statistic	α	δ	μ	λ_s	λ_b	PIN
$\alpha = 0.25$								
	GWJ-LK	MSE	0.0159	0.0827	275444.3	33347.4	34161.9	0.0032
		MAE	0.1041	0.2664	363.9	111.7	112.3	0.0538
	GWJ-EHO	MSE	0.0289	0.0549	386121.7	879044.5	883760.9	0.0022
		MAE	0.1129	0.1905	449.5	578	580.1	0.0377
	YZ-LK	MSE	0.0171	0.1006	287970.9	38682.5	39069.9	0.0054
		MAE	0.1084	0.2977	372.7	120.4	120.8	0.0696
	YZ-EHO	MSE	0.4657	0.1805	453438.0	1828762.2	3746100.5	0.0212
		MAE	0.6623	0.4161	512.0	1132.2	1408.7	0.0945
	BayesPIN	MSE	0.0180	0.0742	241163.0	37502.1	34876.2	0.0036
		MAE	0.1128	0.2528	344.9	117.7	112.4	0.0573
$\alpha = 0.5$								
	GWJ-LK	MSE	0.0216	0.0379	441256.7	14997.4	15408.1	0.0008
		MAE	0.1275	0.1685	507.9	82.2	82.1	0.0241
	GWJ-EHO	MSE	0.0262	0.0577	380058.8	1088283.2	1379179.8	0.0042
		MAE	0.1340	0.1944	445.2	660.5	727.9	0.0542
	YZ-LK	MSE	0.0213	0.0417	444504.2	16086.0	16313.1	0.0016
		MAE	0.1261	0.1782	509.6	84.8	84.4	0.0350
	YZ-EHO	MSE	0.1978	0.1751	495265.1	2091915.0	3114217.3	0.0119
		MAE	0.4353	0.4068	523.0	1193.1	1184.0	0.0789
	BayesPIN	MSE	0.0190	0.0323	418192.9	14669.1	14798.4	0.0008
		MAE	0.1192	0.1556	495.9	79.9	80.2	0.0252
$\alpha = 0.75$								
	GWJ-LK	MSE	0.1066	0.0206	519382.4	8547.9	8313.9	0.0003
		MAE	0.3167	0.1177	564.0	63.8	62.8	0.0134
	GWJ-EHO	MSE	0.0470	0.0663	362286.6	1404685.0	2391455.1	0.0080
		MAE	0.1775	0.2124	431.4	764.6	958.6	0.0748
	YZ-LK	MSE	0.1065	0.0210	521833.9	8663.4	8452.1	0.0005
		MAE	0.3164	0.1191	564.2	64.4	63.3	0.0171
	YZ-EHO	MSE	0.0483	0.1738	608385.0	2041920.5	2932617.8	0.0111
		MAE	0.2126	0.4048	559.7	1179.8	1054.0	0.0813
	BayesPIN	MSE	0.1027	0.0174	502205.7	8188.8	7987.9	0.0003
		MAE	0.3114	0.1086	556.3	62.4	61.5	0.0125

Table 10: Comparison of parameter estimation methods for varying α and fixed $\delta = 0.5$ for UIG Scale

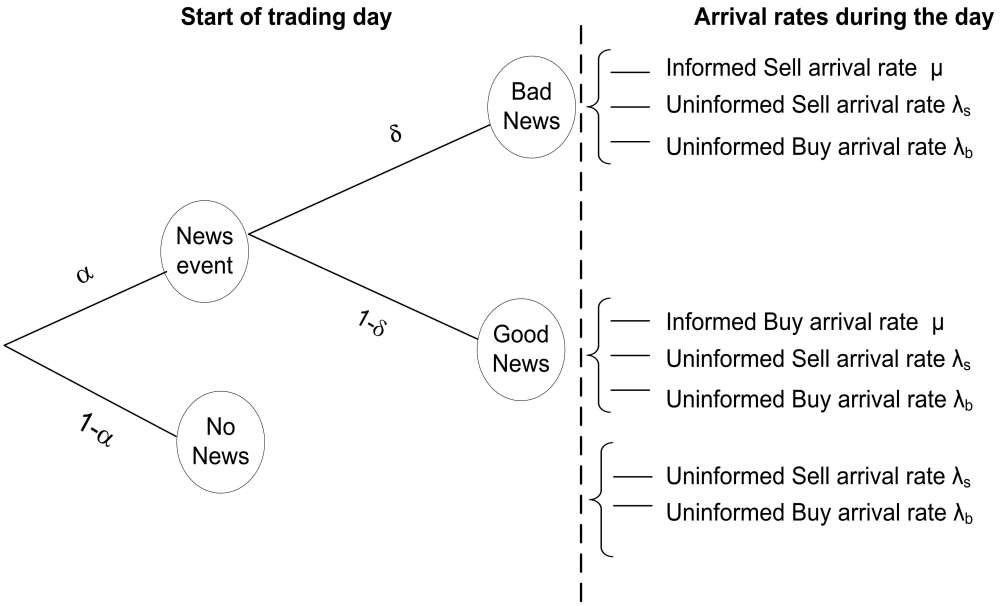


Fig. 1. Tree diagram for Easley et al. (2002) model

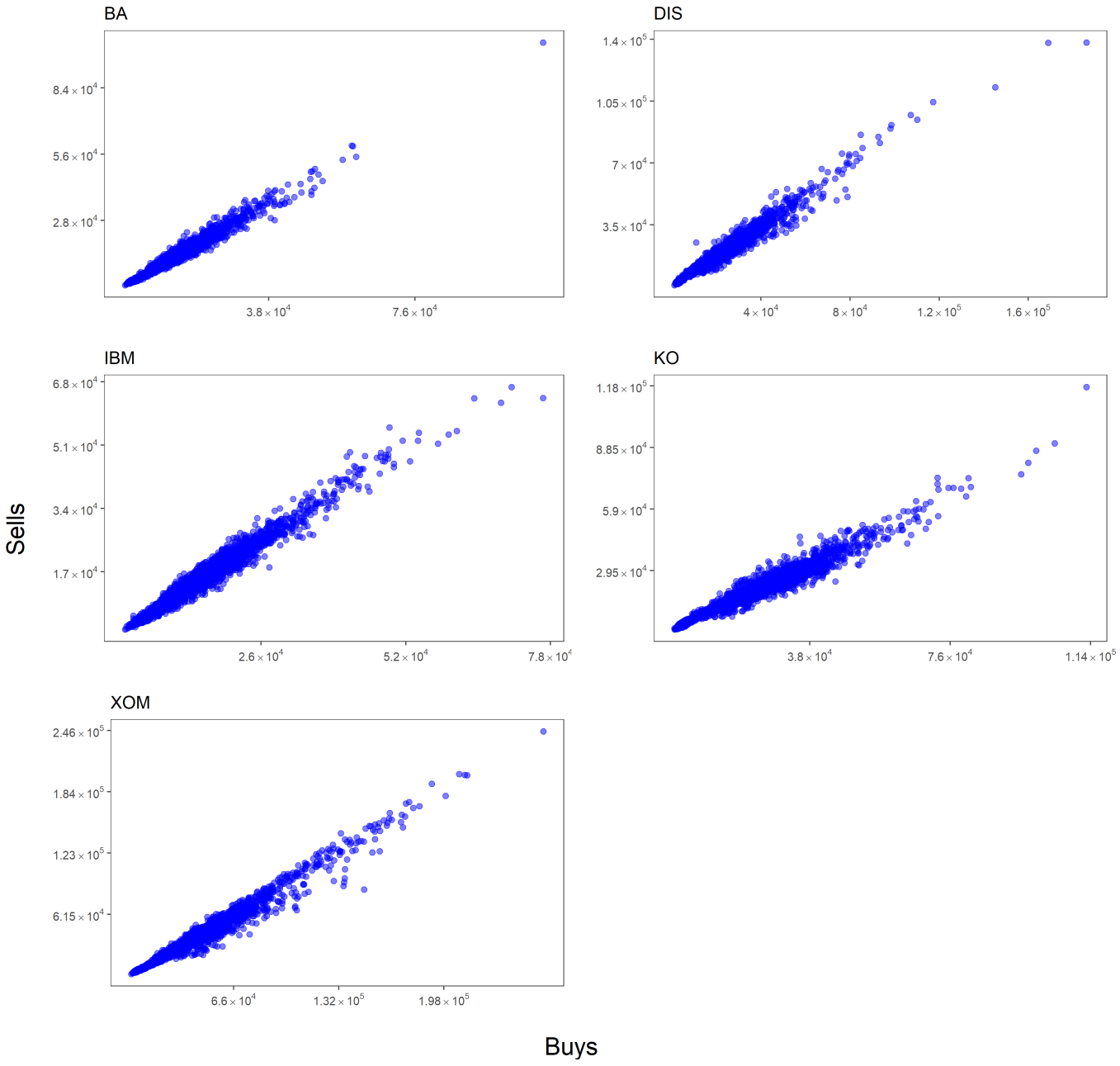


Fig. 2. Scatter plot of daily buy and sell trades

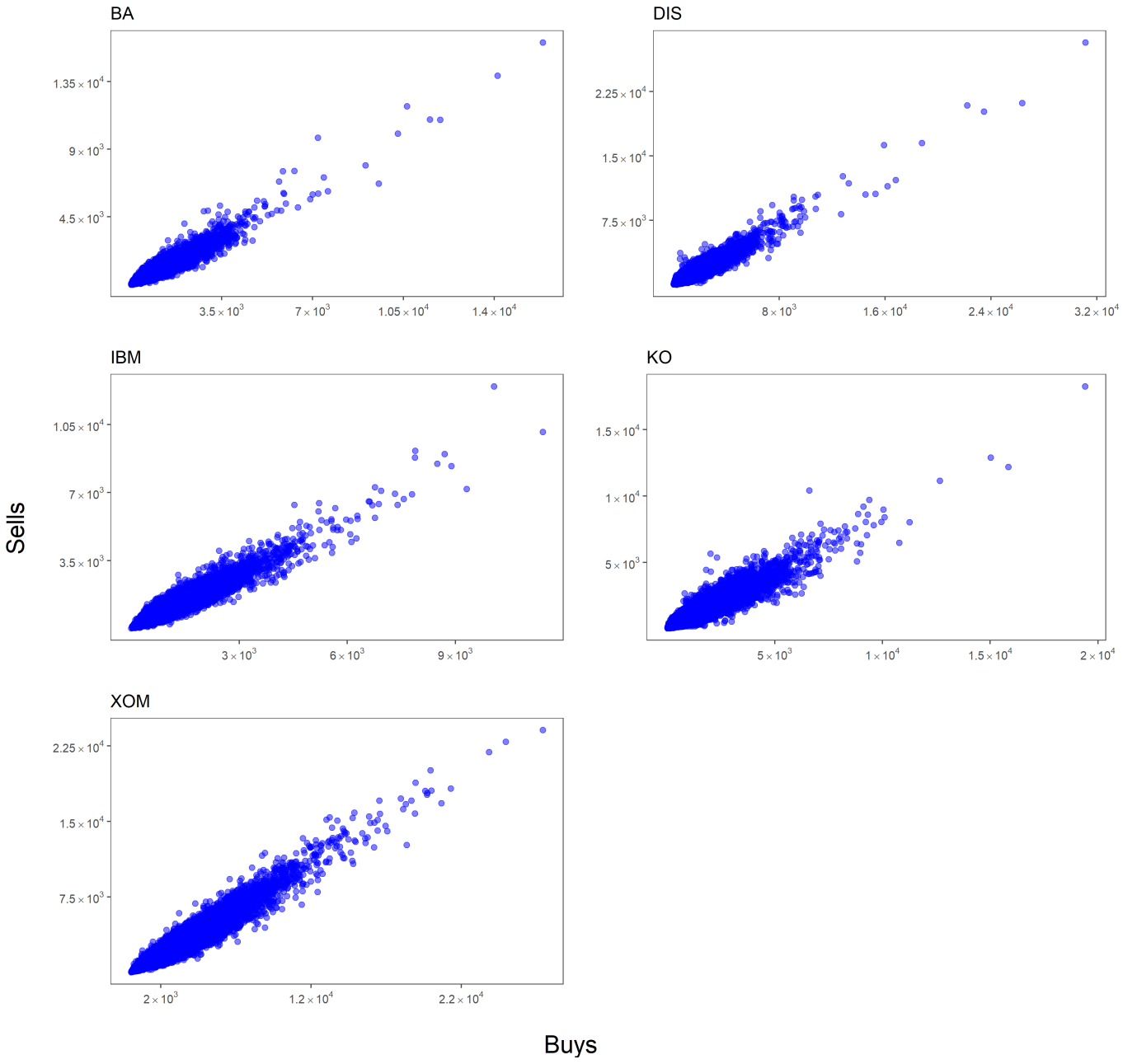
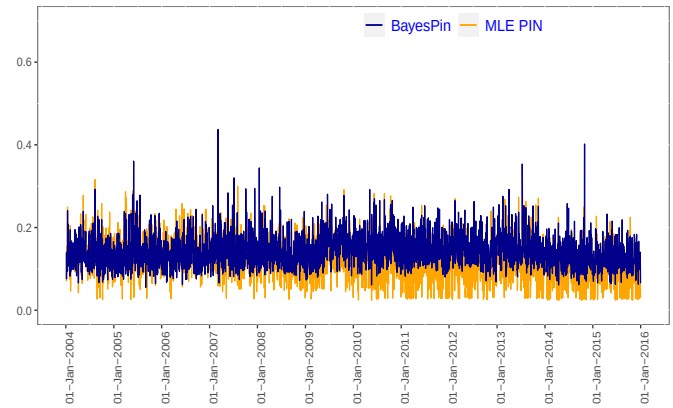
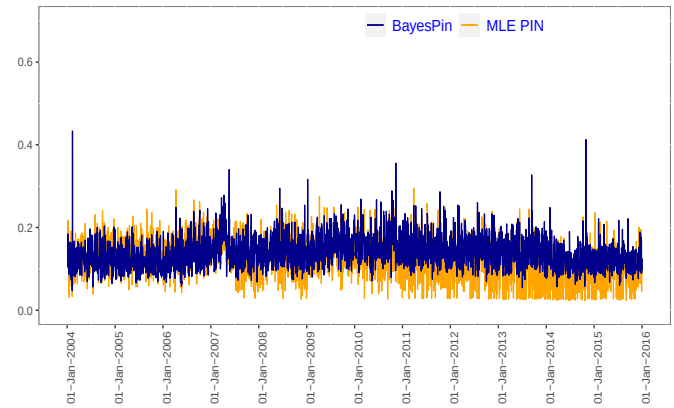


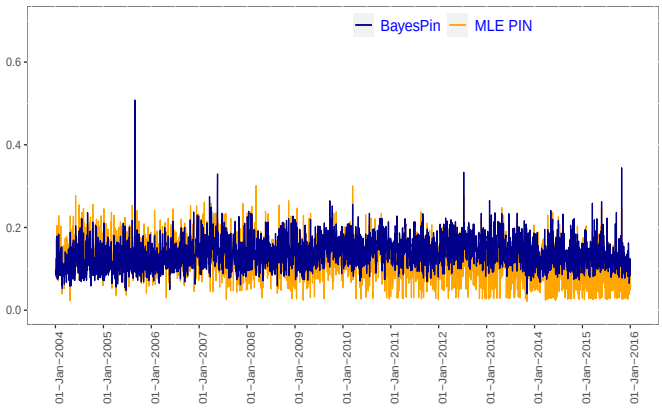
Fig. 3. Scatter plot of 15−minute sampled buy and sell trades



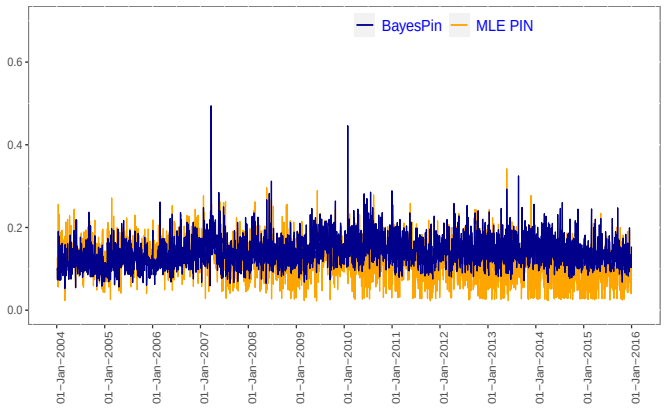
(a) BA



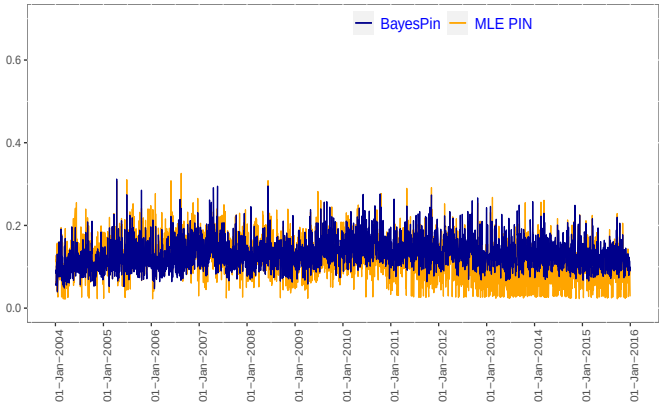
(b) DIS



(c) IBM



(d) KO



(e) XOM

Fig. 4. Daily PIN estimates from intraday data using the InfoTrad package and the Bayesian algorithm

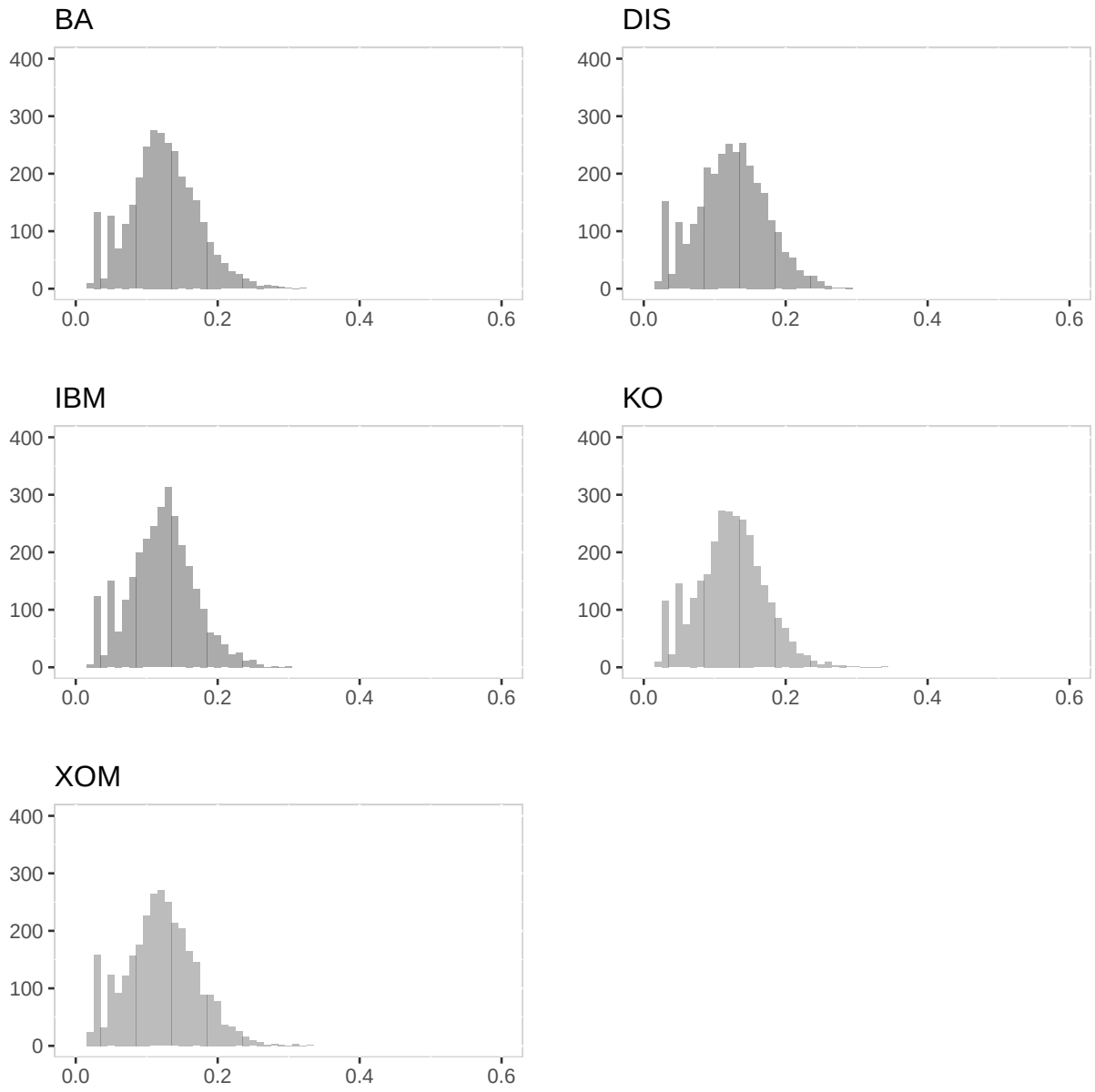


Fig. 5. Histogram of Daily PIN estimates from intraday data using the InfoTrad package

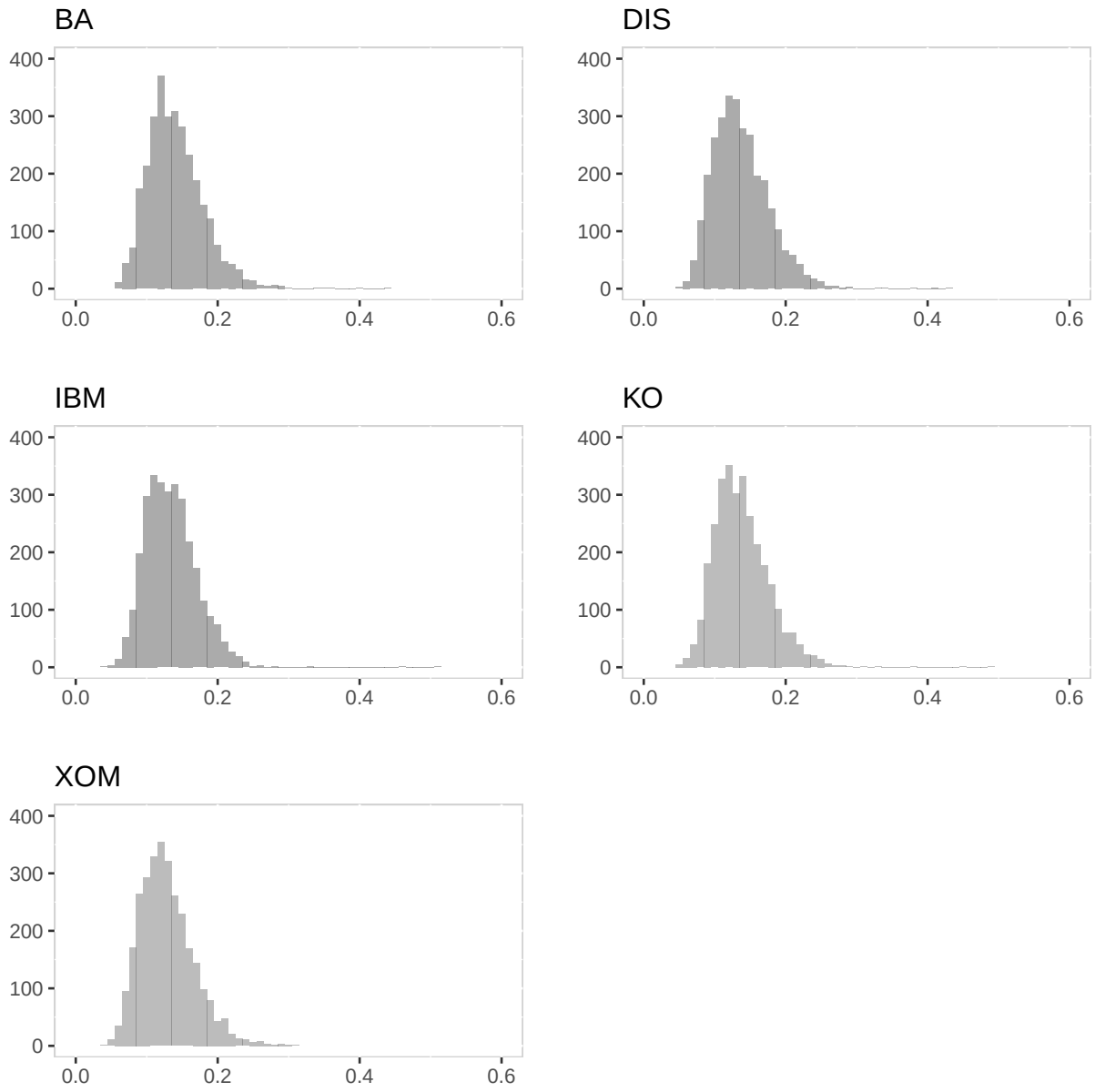
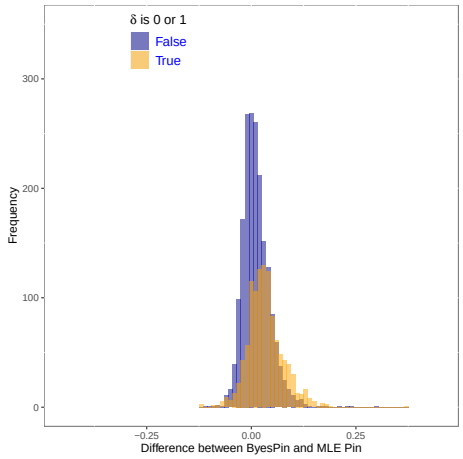
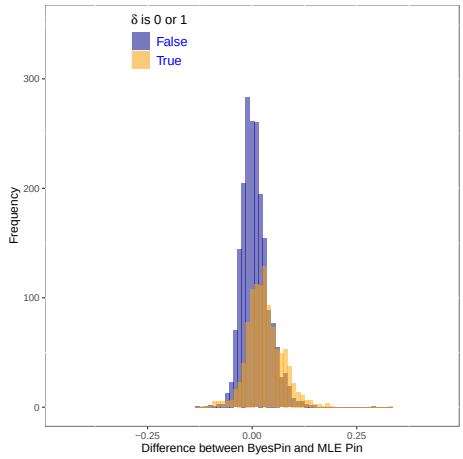


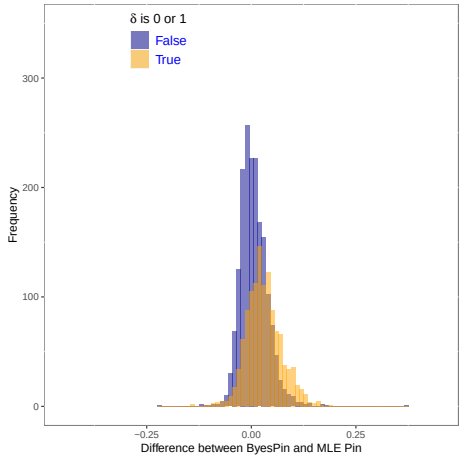
Fig. 6. Histogram of Daily PIN estimates from intraday data using the Bayesian algorithm



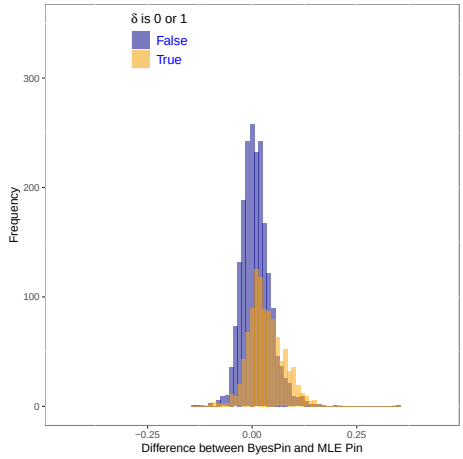
(a) BA



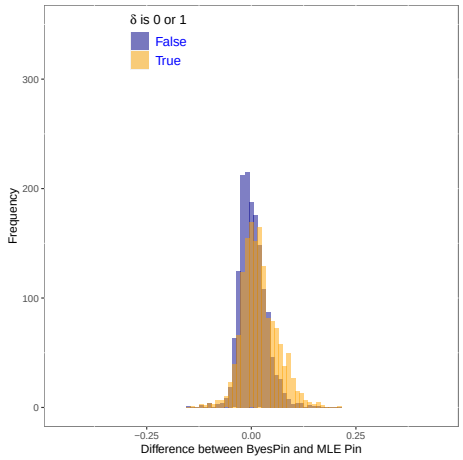
(b) DIS



(c) IBM



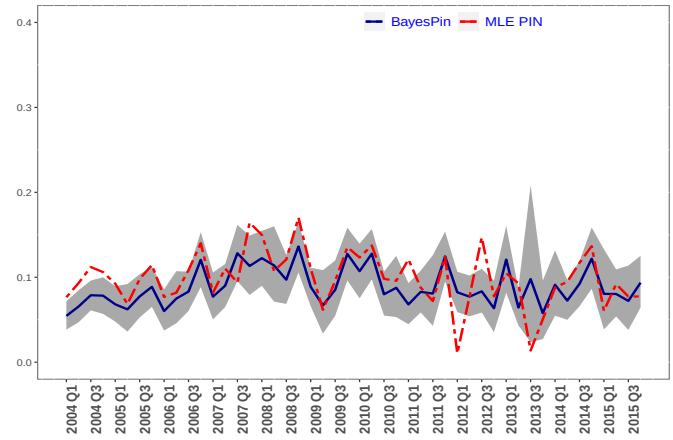
(d) KO



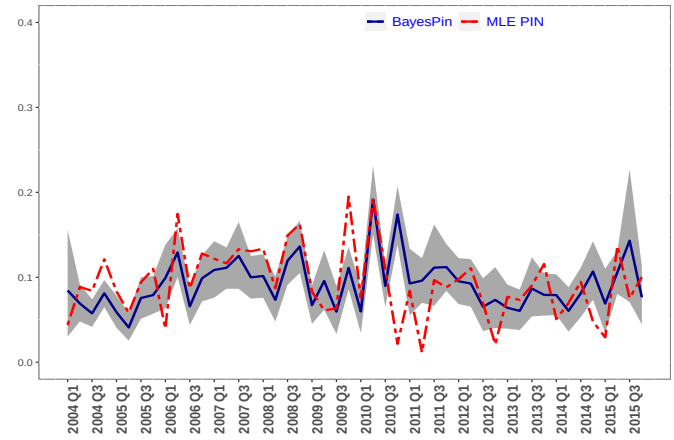
(e) XOM

Fig. 7. Histograms of the differences between daily Bayesian and MLE Pin estimates.
Note: The darker shaded areas (blue) represent the regular cases while the lighter shaded areas (gold) represent corner solutions in MLE.

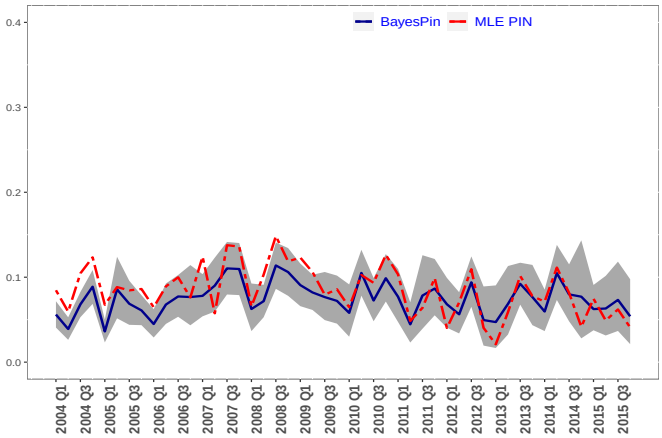
https://www.overleaf.com/project/5eb917c293f026000132dafa



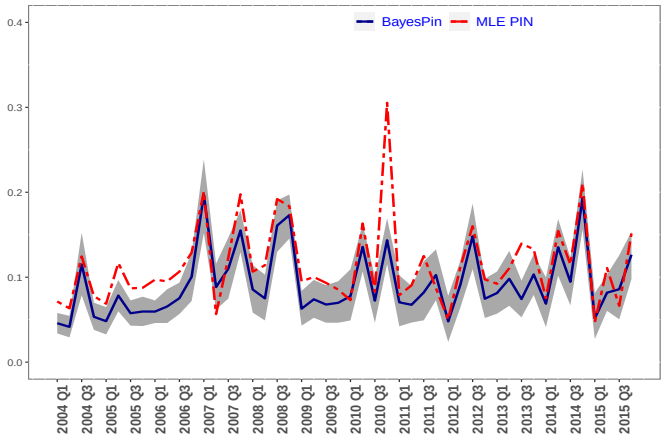
(a) BA



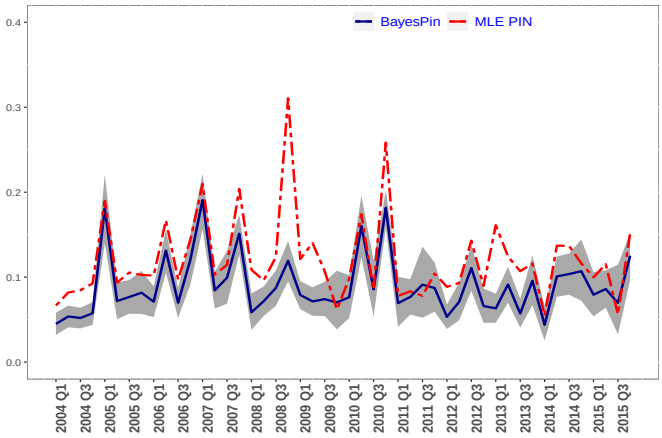
(b) DIS



(c) IBM



(d) KO



(e) XOM

Fig. 8. Quarterly BayesPin estimates from daily aggregated buys and sells with credible intervals.

Note: The solid lines (blue) represent the BayesPin estimates while the dashed lines (red) represent MLE PIN estimates. The grey shaded region is the credible interval for the BayesPin estimates.

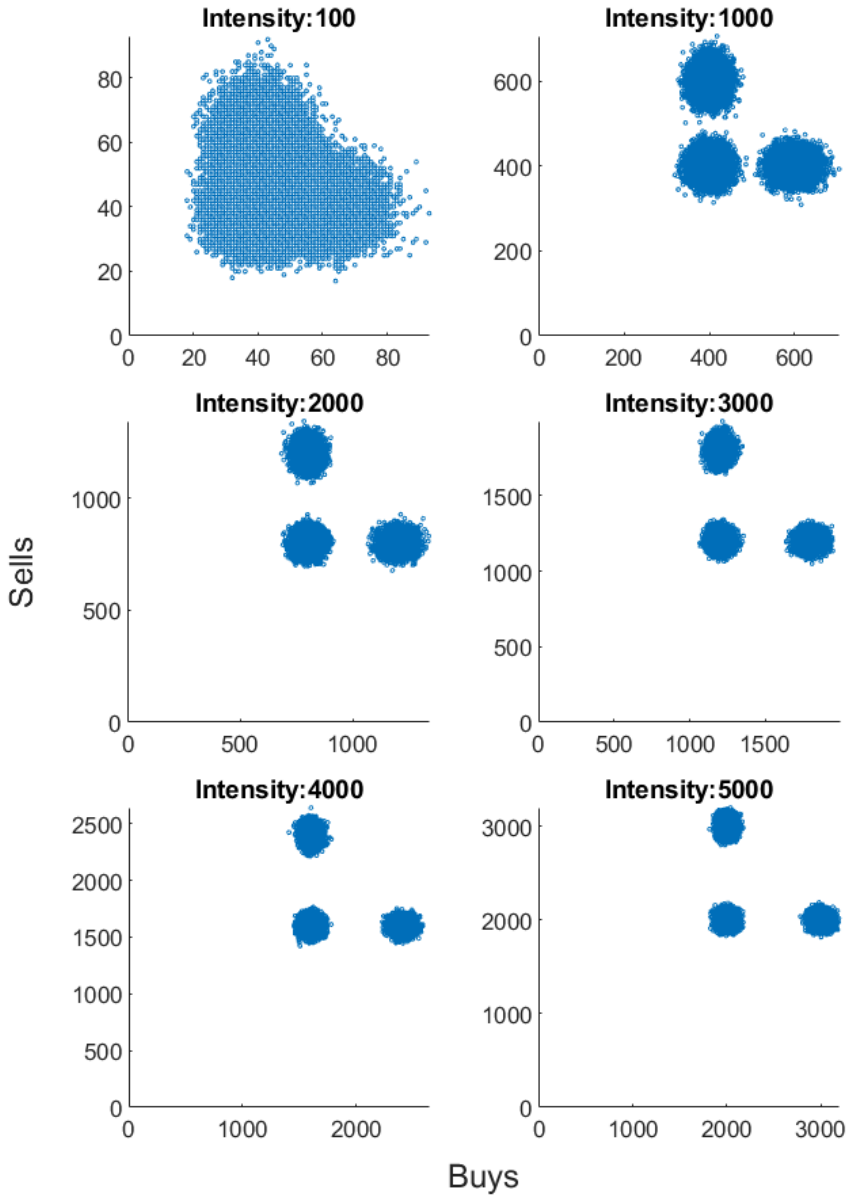


Fig. 9. Scatter plot of simulated buy and sell trades at varying trade intensities, with $\alpha = 0.5$, $\delta = 0.5$, $\mu = 0.2$, and $\lambda_b = \lambda_s = 0.4$, so that PIN$= 0.11$. $60,000$ pairs are generated at each intensity.

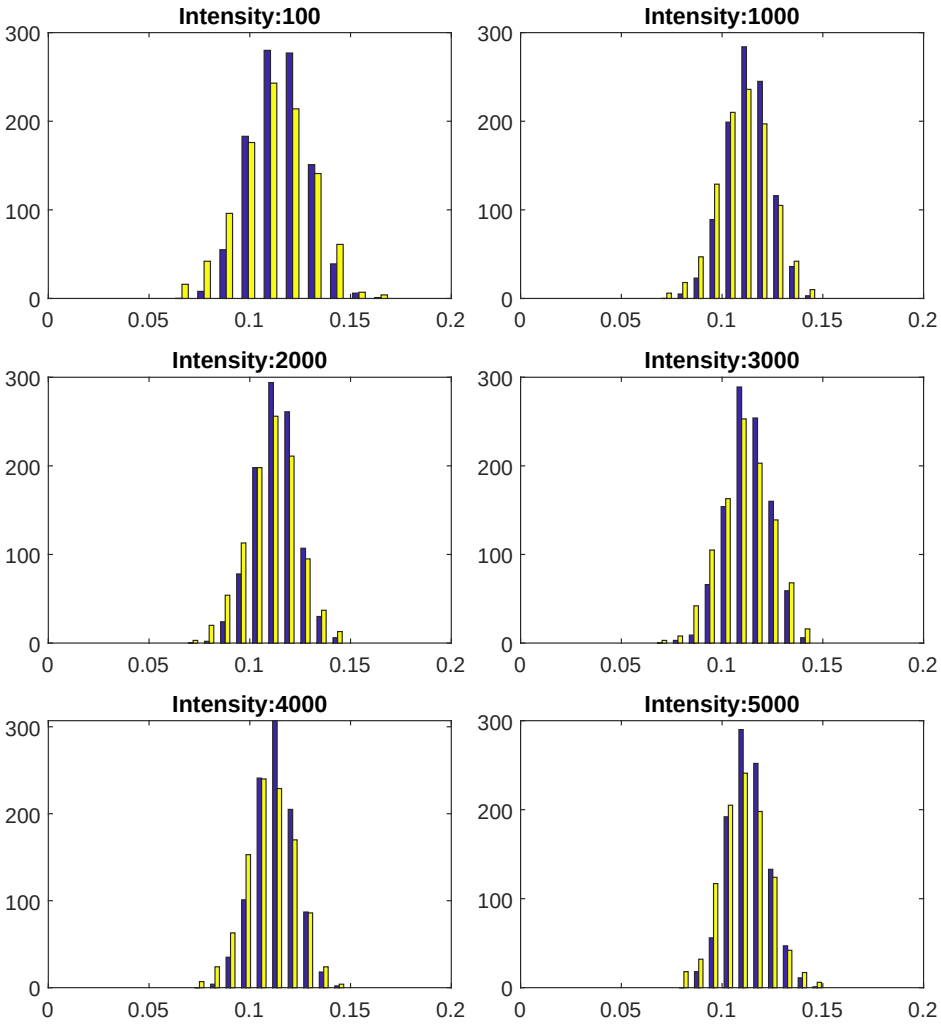


Fig. 10. Histogram of estimates of PIN at 6 different trade intensities. The blue (darker) columns represent the BayesPIN algorithm and the yellow (lighter) columns the MLE estimates with GWJ algorithm and LK factorization. The true value in each case is 0.11.

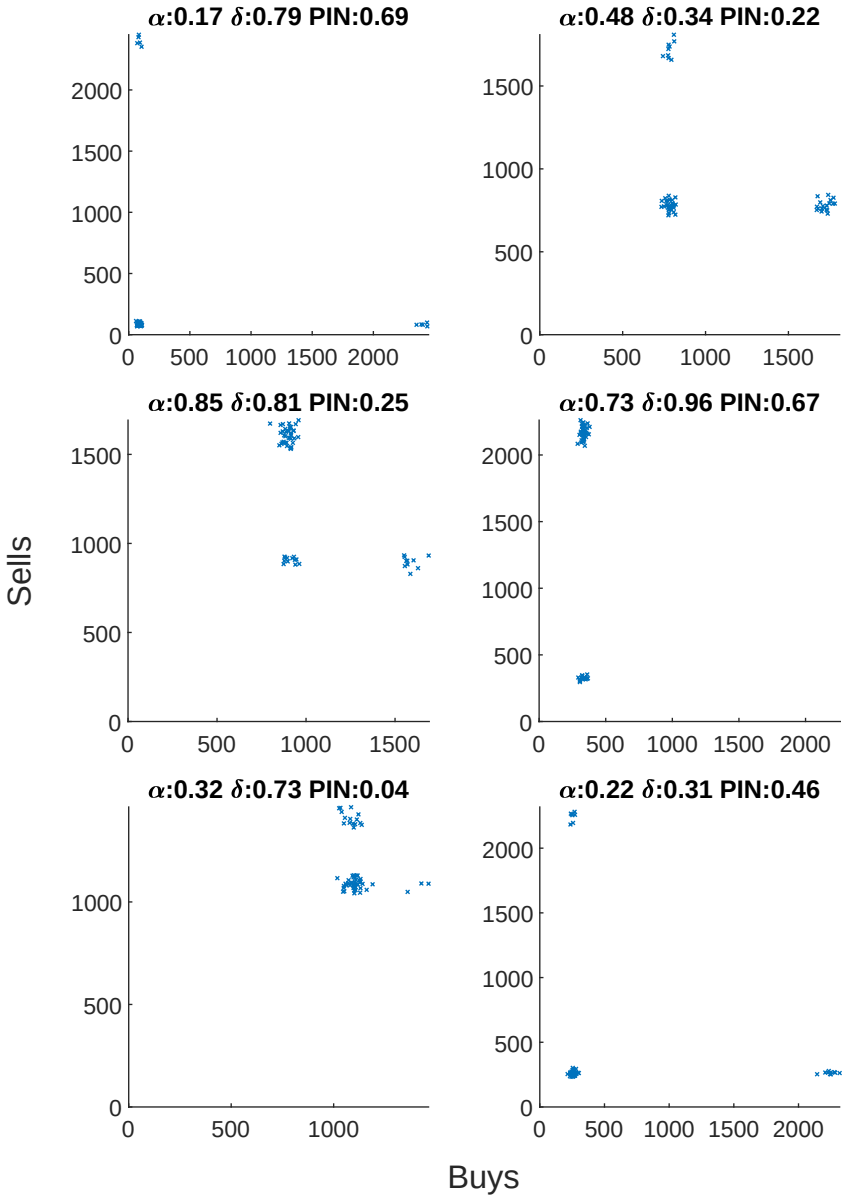


Fig. 11. Scatter plot of randomly selected sets of 60 simulated buy and sell trades where parameters are randomly drawn, at trade intensity 2500

Note: Each case represents different combinations of values of the parameters, shown above the panels.

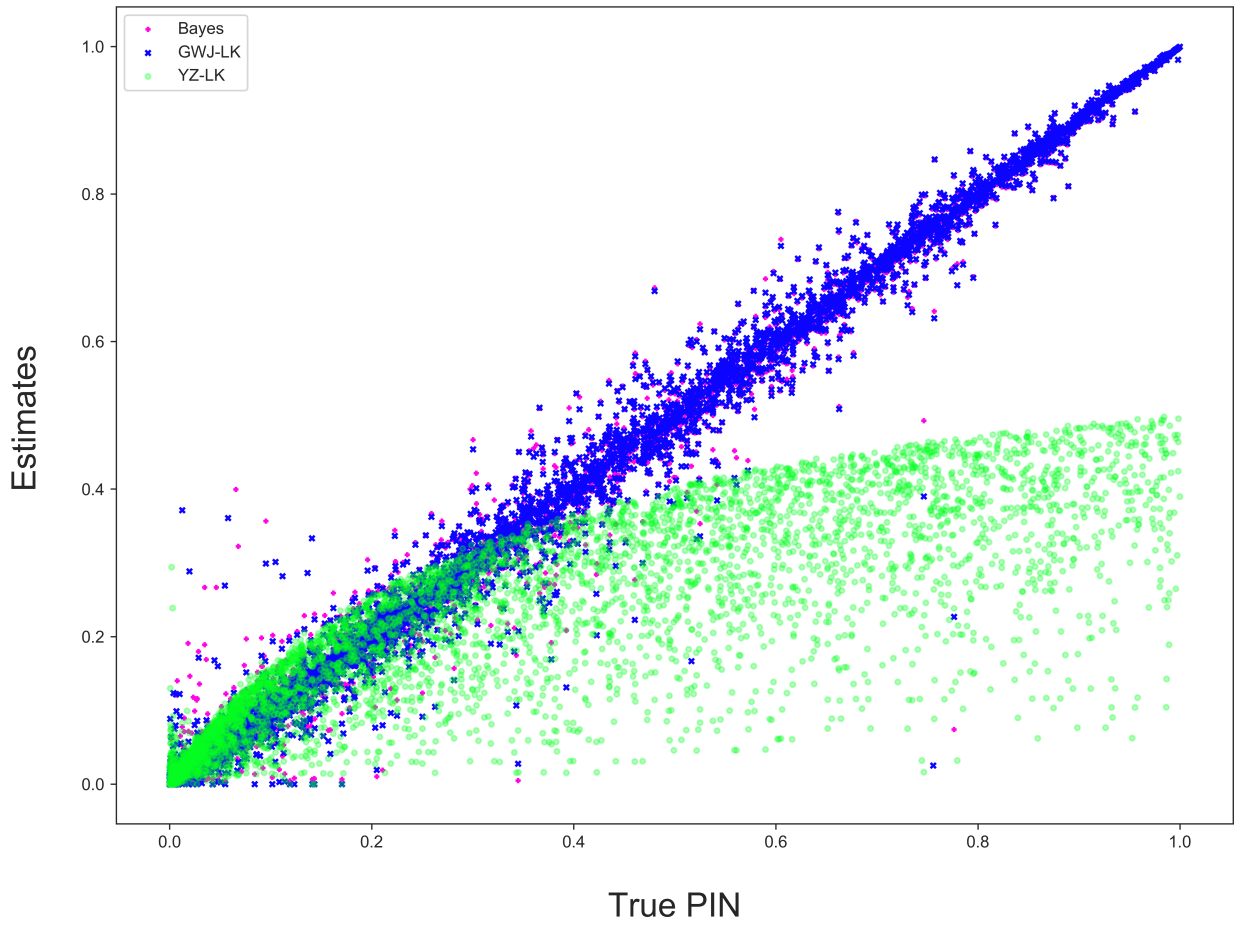


Fig. 12. Scatter of three sets of PIN estimates: for the Bayesian algorithm and two MLE approaches.

Note: We simulated $5,000$ sets of 60 observations of buy and sell trades by drawing uniform random values corresponding to α, δ, and $\mu/2500$. The remaining trade intensity $2500(1-\mu)$ was allocated equally between uninformed buy and sell intensitiesλ_b and λ_s.

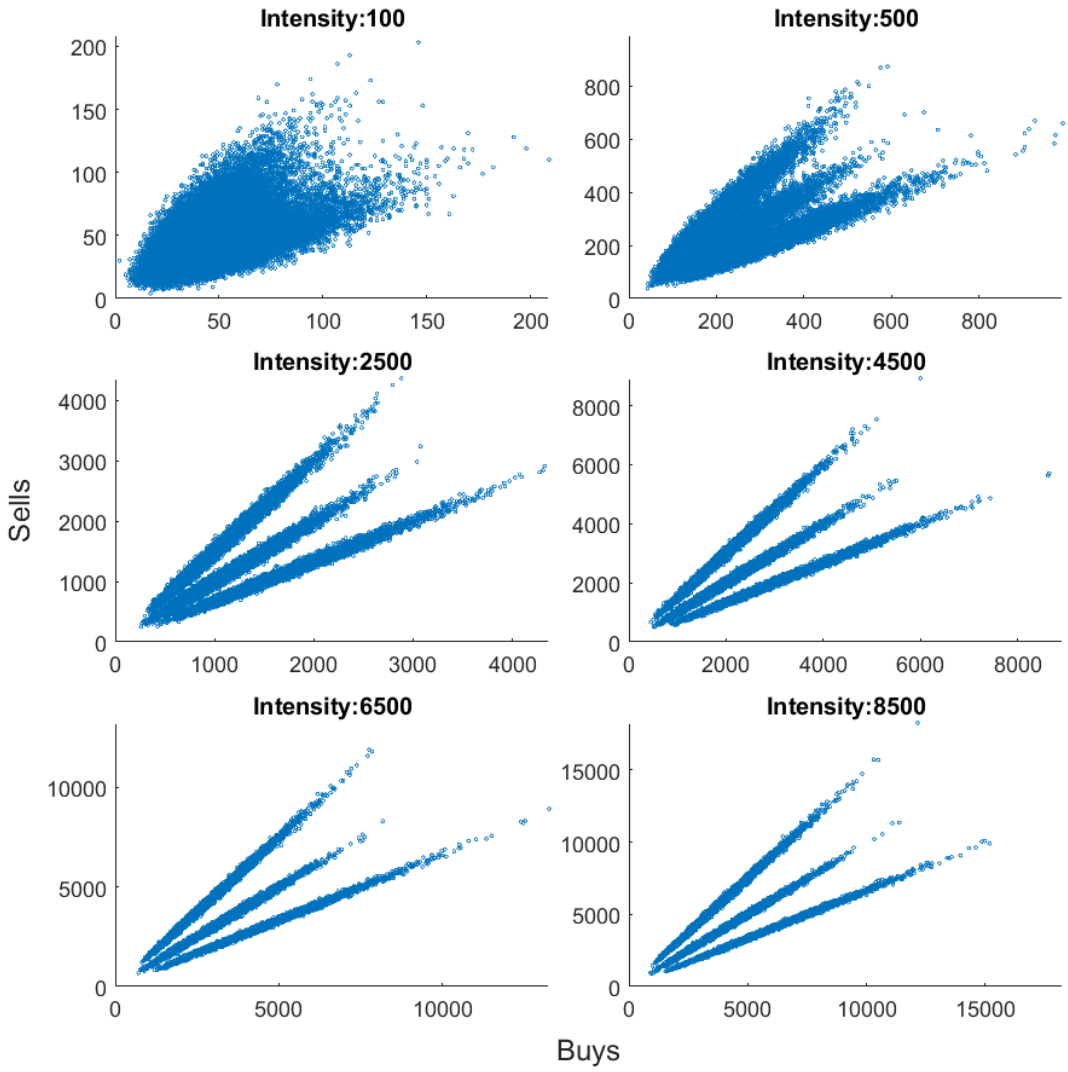


Fig. 13. Scatter plot of simulated buy and sell trades at varying trade intensities scaled by a random Unit Inverse Gaussian scaling factor, with $\alpha = 0.5$, $\delta = 0.5$, $\mu = 0.2$, $\lambda_b = \lambda_s = 0.4$, and $\psi = 9$, so that PIN$= 0.11$. $60,000$ pairs are generated at each intensity.