Journal of Banking & Finance 43 (2014) 137–149

Contents lists available at ScienceDirect

Journal of Banking & Finance

journal homepage: www.elsevier.com/locate/jbf

Quality of PIN estimates and the PIN-return relationship

Yuxing Yan [a,1], Shaojun Zhang [b,*]

[a] Department of Economics and Finance, Canisius College, 2001 Main Street, Buffalo, NY 14208, United States
[b] School of Accounting and Finance, Faculty of Business, Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 31 October 2012
Accepted 5 March 2014
Available online 20 March 2014

JEL classification:
C13
C61
G12
G14

Keywords:
Probability of informed trading
PIN-return relationship
Quality of PIN estimates

ABSTRACT

This paper provides new evidence concerning the probability of informed trading (PIN) and the PIN-return relationship. We take measures to overcome known estimation biases and improve the quality of quarterly PIN estimates. We use the average of a firm's PIN estimates in four consecutive quarters to smooth out the effect of seasonal variation in trading activities. We find that when high-quality PIN estimates are used, the Fama–MacBeth cross-sectional regressions show stronger evidence for the positive PIN-return relationship than documented in the prior literature. This finding is robust to controls for the January, liquidity, and momentum effects.

1. Introduction

A vibrant finance and accounting literature attempts to understand whether and how information asymmetry between investors influences asset prices in financial markets. In an influential paper, Easley et al. (2002) use the probability of informed trading (PIN) measure to quantify the degree of information asymmetry and document a significant positive relationship between the PIN measure and stock returns between 1983 and 1998. The positive PIN-return relationship has been widely cited in many studies.[2]

However, some researchers doubt the existence of a significant cross-sectional PIN-return relationship for several reasons. First, the relationship is ambiguous in theory. Easley and O'Hara (2004) demonstrate that, in a finite economy, uninformed investors demand a risk premium for holding stocks of greater information asymmetry because they cannot diversify the risk of trading against informed investors. On the other hand, Hughes et al. (2007) show that, if taking a large economy limit and holding the total information constant, greater information asymmetry in the aggregate information environment leads to higher market cost of capital, but firm-specific information characteristics do not affect individual firm expected return after controlling for systematic factors.

Second, the PIN measure must be estimated via a maximum likelihood approach using the daily number of buyer-initiated and seller-initiated trades. A few studies point out potential biases that arise in the estimation of PIN. Boehmer et al. (2007) find that misclassification of buyer-initiated and seller-initiated trades leads to a downward bias in PIN estimates. Lin and Ke (2011) prove that the mathematical transformation used by Easley et al. (2002, 2010) to simplify the joint likelihood function generates a bias because of computer floating-point exception. Yan and Zhang (2012) show that boundary solutions can be another source of bias. Duarte and Young (2009) argue that the microstructure model used by Easley et al. (2002) to describe the trading process does not distinguish information asymmetry from illiquidity.[3]

Third, several studies find that empirical evidence on the PIN-return relationship is not robust. Mohanram and Rajgopal (2009) report that the PIN-return relationship is significantly positive only in the period 1984–1988. It is insignificant in two periods, 1989–1993 and 1994–1998, and is negative in the period 1999–2002. Kang (2011) presents evidence for a January effect, that is, PIN and

* Corresponding author. Tel.: +852 3400 3458; fax: +852 2330 9845.
E-mail addresses: yany@canisius.edu (Y. Yan), afszhang@polyu.edu.hk (S. Zhang).
[1] Tel.: +1 (716) 888 2604.

[2] See, e.g., Francis et al. (2004, 2005), Odders-White and Ready (2006), Chen et al. (2007), Duarte et al. (2008), Brockman and Yan (2009), and Chen and Zhao (2012).

[3] Duarte and Young (2009) use an extended model to propose another measure of information asymmetry and call it the adjusted PIN. Because many studies have used the same PIN measure as Easley et al. (2002), our focus is to examine whether the quality of PIN estimates affects the PIN-return relationship.

http://dx.doi.org/10.1016/j.jbankfin.2014.03.006
0378-4266/

stock returns are negatively related in January, but positively related in other months.

This paper provides new evidence on the PIN-return relationship using high-quality PIN estimates for over 170,000 stock-quarter pairs in the 22 years between 1983 and 2004.[4] We apply several methods that are developed to overcome known estimation biases in order to improve the quality of quarterly PIN estimates. We implement the Fama and MacBeth (1973) regression methodology over the period of 276 months between April 1983 and March 2005. It is evident that the PIN-return relationship is significantly positive over the whole period. Moreover, we examine the same four sub-periods, consistent with Mohanram and Rajgopal (2009): 1984–1988, 1989–1993, 1994–1998, and 1999–2002. Contrary to their observation that the coefficient of PIN is only significant in the earliest sub-period 1984–1988, we find that it is significantly positive in both 1984–1988 and 1994–1998, and its magnitude is larger than that found by Mohanram and Rajgopal (2009). This finding is robust after we adjust for time-varying precision in monthly regression estimates with the Litzenberger and Ramaswamy (1979) weighted least-square method, exclude January observations from the analysis, and control for the liquidity and momentum effects in the cross-sectional regressions.

We contribute to the literature in three ways in addition to documenting new evidence on the PIN-return relationship. First, the Lee and Ready (1991) classification algorithm with a five-second time adjustment has been commonly applied to estimate PIN in previous studies (see, e.g., Easley et al., 2002, 2010; Duarte and Young, 2009; Brown et al., 2004; Yan and Zhang, 2012). This paper is the first to document empirical evidence that the five-second time adjustment causes a systematic bias in PIN estimates for a substantial number of actively traded stocks in the years after 2000.

Second, Boehmer et al. (2007) show that trade misclassification can result in a *downward* bias in PIN estimates. They study a microstructure model in which the arrival rates of buy and sell trades are assumed to be the same. Previous studies often use another model that allows the two arrival rates to be different (see, e.g., Easley et al., 2002, 2010; Duarte et al., 2008; Brown et al., 2004; Yan and Zhang, 2012). Our simulations demonstrate that trade misclassification may result in an *upward* bias under both models.

At last, we observe a distinct seasonal pattern, that is, the PIN estimates on average tend to decrease in the first quarter of a year relative to the previous quarter. Our preliminary analysis suggests that this seasonal pattern is related to tax-loss selling activities at year end. This finding prompts us to use the average of a firm's PIN estimates in four consecutive quarters to smooth out the effect of seasonal variation in trading activities.

The remainder of this paper is organized as follows. Section 2 reviews the estimation of PIN and trade classification. Section 3 compares the two sets of quarterly PIN estimates that are obtained with two different trade classification methods, reports our findings from a simulation study, and presents a preliminary analysis of seasonal variation in the probability of informed trading. Section 4 reports empirical evidence on the PIN-return relationship and its robustness. Section 5 concludes the paper.

2. Trade classification and the estimation of PIN

2.1. The estimation of PIN

The PIN measure of information asymmetry is derived from the market microstructure model proposed in Easley and O'Hara (1992) and Easley et al. (1997). Mathematically, the model specifies that on any day i, the likelihood of observing the number of buy trades B_i and the number of sell trades S_i is represented by

$$L(\theta|B_i,S_i)=\alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^{B_i}}{B_i!}e^{-\varepsilon_s}\frac{\varepsilon_s^{S_i}}{S_i!}+\alpha\delta e^{-\varepsilon_b}\times\frac{\varepsilon_b^{B_i}}{B_i!}e^{-(\mu+\varepsilon_s)}\frac{(\mu+\varepsilon_s)^{S_i}}{S_i!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^{B_i}}{B_i!}e^{-\varepsilon_s}\frac{\varepsilon_s^{S_i}}{S_i!} \quad (1)$$

where $\theta=(\alpha,\delta,\mu,\varepsilon_b,\varepsilon_s)$ represents five structural parameters that describe the trading process in each day. Specifically, α denotes the probability that an information event occurs. If an information event occurs, it can be bad news with the probability δ or good news with the probability $1-\delta$, and informed traders who know the quality of new information submit orders at the daily arrival rate μ. Informed traders would buy at the rate μ if it is good news, and sell at the same rate μ if it is bad news. No matter whether an information event occurs or not, uninformed traders submit buy orders at the daily arrival rate ε_b and sell orders at the daily arrival rate ε_s.

Assuming independence between days, the joint likelihood of observing a series of daily buys and sells over trading days $i=1,\ldots,I$ is the product of the daily likelihoods,

$$L(\theta|M)=\prod_{i=1}^{I}L(\theta|B_i,S_i) \quad (2)$$

where $M=((B_1,S_1),\ldots,(B_I,S_I))$ represents the data set. The PIN measure of information asymmetry is defined as

$$\mathrm{PIN}=\frac{\alpha\mu}{\alpha\mu+\varepsilon_B+\varepsilon_S}. \quad (3)$$

Intuitively, PIN equals the fraction of trades in a day that arise from informed trading.

Maximizing the joint likelihood in Eq. (2) over the parameters in θ produces the maximum likelihood estimates of these structural parameters. There is no closed form solution to this maximization problem. A numerical maximization technique must be used to obtain a solution. Easley et al. (2010) use the following factorization of the joint likelihood function to facilitate numerical maximization

$$L((B_i,S_i)_{i=1}^{I}|\theta)=\sum_{i=1}^{I}[-\varepsilon_b-\varepsilon_s+M_i(\ln x_b+\ln x_s)+B_i\ln(\mu+\varepsilon_b)+S_i\ln(\mu+\varepsilon_s)]+\sum_{i=1}^{I}\ln[\alpha(1-\delta)e^{-\mu}x_x^{S_i-M_i}x_b^{-M_i}+\alpha\delta e^{-\mu}x_b^{B_i-M_i}x_s^{-M_i}+(1-\alpha)x_s^{S_i-M_i}x_b^{B_i-M_i}] \quad (4)$$

where $M_i=\min(B_i,S_i)+\max(B_i,S_i)/2$, $x_s=\frac{\varepsilon_s}{\mu+\varepsilon_s}$, and $x_b=\frac{\varepsilon_b}{\mu+\varepsilon_b}$. Here, $\min(B_i,S_i)$ represents the smaller of B_i and S_i, and $\max(B_i,S_i)$ represents the larger one.

Lin and Ke (2011) point out that when the above factorized likelihood function is used, floating-point exception in computer software narrows the set of feasible solutions, which causes a downward bias in the estimate of PIN. In order to avoid the influence of floating-point exception, they recommend the following factorization of the joint likelihood function

$$L((B_i,S_i)_{i=1}^{I}|\theta)=\sum_{i=1}^{I}[-\varepsilon_b-\varepsilon_s+B_i\ln(\mu+\varepsilon_b)+S_i\ln(\mu+\varepsilon_s)+e_{\mathrm{max}i}]+\sum_{i=1}^{I}\ln[\alpha(1-\delta)\exp(e_{1i}-e_{\mathrm{max}i})+\alpha\delta\exp(e_{2i}-e_{\mathrm{max}i})+(1-\alpha)\exp(e_{3i}-e_{\mathrm{max}i})] \quad (5)$$

[4] We do not extend the estimates beyond 2004 for two reasons. First, the implementation of Regulation NMS in 2005 had substantial impact on trading activities, which increases the difficulty in reliably classifying buyer-initiated and seller-initiated trades. Second, it is financially costly for us to gain access to the intraday trade and quote data and to use a powerful computing platform. Our PIN estimates are available upon request.

where $e_{1i} = -\mu - S_i \ln(1 + \mu/\varepsilon_s)$, $e_{2i} = -\mu - B_i \ln(1 + \mu/\varepsilon_b)$, $e_{3i} = -B_i \ln(1 + \mu/\varepsilon_b) - S_i \ln(1 + \mu/\varepsilon_s)$, and $e_{maxi} = \max(e_{1i}, e_{2i}, e_{3i})$.

Numerical methods such as the modified Newton–Raphson method are used to find the maximizing parameters of the factorized likelihood function in both Eqs. (4) and (5). Yan and Zhang (2012) argue that numerical solutions may erroneously fall on the boundary of the parameter space, which causes a bias in PIN estimate. In order to avoid boundary solutions, they use the method of moment to choose 125 sets of initial values that can be used to explore the parameter space efficiently. More specifically, the initial values for the five parameters in Eq. (1) are given by

$$\alpha^0 = \alpha_i, \delta^0 = \delta_j, \varepsilon_b^0 = \gamma_k \cdot \overline{B}, \mu^0 = \frac{\overline{B} - \varepsilon_b^0}{\alpha^0 \cdot (1 - \delta^0)}, \quad \text{and}$$
$$\varepsilon_s^0 = \overline{S} - \alpha^0 \cdot \delta^0 \cdot \mu^0 \tag{6}$$

where the three variables α_i, δ_j, and γ_k take values from the five fractions (0.1, 0.3, 0.5, 0.7, 0.9), one at a time. The combinations of α_i, δ_j, and γ_k yield 125 sets of initial values. The combinations that have a negative value for ε_s^0 are eliminated.

2.2. Trade classification

The estimation of PIN requires both the number of buyer-initiated trades and the number of seller-initiated trades in each day. However, the majority of publicly accessible trade and quote databases do not identify whether a reported trade is initiated by a buyer or a seller. Researchers must classify trades with certain rules.

The *tick* rule classifies a trade as a buy (sell) if the trade price is higher (lower) than the previous trade. If the current and previous trade prices are the same, the trade is classified by the next previous trade. The *quote* rule classifies a trade as a buy (sell) if the trade price is closer to the prevailing ask (bid) quote. This rule is unable to classify trades at the midpoint between bid and ask quotes. Lee and Ready (1991) propose an algorithm that combines the tick and quote rules: the tick rule is applied to trades that occur at the midpoint of the prevailing bid and ask quotes, while the quote rule is applied to classify any trade that takes place above (below) the quote midpoint as a buy (sell).

Ellis, Michaely, and O'Hara (EMO, 2000) propose another algorithm that outperforms the LR algorithm in their study of Nasdaq stocks. According to their algorithm, all trades executed at the ask quote are categorized as buys, all trades executed at the bid quote are categorized as sells, and all the other trades are categorized by the tick rule. Peterson and Sirri (2003) examine the accuracy of these two algorithms for NYSE/AMEX stocks and find that the EMO algorithm, combined with using contemporaneous quotes, provides the least amount of bias in execution cost measurement.

The reported time of trades and quotes in the publicly available database has a significant impact on the accuracy of trade classification. Lee and Ready (1991) suggest that because of the time delay in reporting trades, the reported trade time should be reduced by 5 s in order to find the prevailing quotes. EMO (2000) find that it works fine not to adjust the reported trade time for Nasdaq stocks. Peterson and Sirri (2003) present evidence in favor of not adjusting the reported trade time for NYSE stocks. Bessembinder (2003) shows that making no allowance for trade reporting delay is optimal when assessing whether trades are buyer and seller initiated, for both Nasdaq and NYSE stocks. Henker and Wang (2006) also suggest that for NYSE stocks, the prevailing quotes should be the ones immediately before the trades in the TAQ database. However, it has been a common practice in the PIN literature that the Lee and Ready classification algorithm with five-second adjustment is used to classify trades (see, e.g., Easley et al. (2002, 2010), Duarte and Young (2009), Brown et al. (2004), Yan and Zhang (2012)). In this paper, we report evidence on whether and how the five-second adjustment for delay in reported trade time influences PIN estimates.

3. The empirical properties of our PIN estimates

3.1. The differences between two sets of estimates

We estimate PIN for the NYSE/AMEX stocks that have data in the ISSM and TAQ databases between January 1, 1983 and December 31, 2004. Stocks in the ISSM and TAQ databases are matched with those in the CRSP database by the historical eight-digit CUSIP. We keep only stocks with the CRSP share code 10 or 11, which means that the closed-end funds, real estate investment trusts, American depository receipts, and foreign stocks are excluded. We estimate PIN for each stock on a quarterly basis and require that the stock has trades and quotes for at least 20 trading days in one quarter.

We use the Lin and Ke (2011) factorized joint likelihood function in Eq. (5). We also apply the Yan and Zhang (2012) algorithm to construct 125 sets of initial values and choose the non-boundary numerical solution that maximizes the likelihood function. Thus, our estimates are free from the biases caused by floating-point exception and boundary solutions.

The estimation of PIN requires both the number of buyer-initiated trades (i.e., buy trades) and the number of seller-initiated trades (i.e., sell trades) in each day. We classify trades as buyer- or seller-initiated using two methods: the Lee and Ready (LR, 1991) algorithm with a five-second time adjustment and the EMO (2000) algorithm without time adjustment. Thus, we obtain two sets of PIN estimates: the LR estimates and the EMO estimates. Fig. 1 compares the two sets of PIN estimates in each quarter from the first quarter of 1983 (1983Q1) to the fourth quarter of 2004 (2004Q4).[5] The number of stocks that have both LR and EMO estimates ranges from 1741 in 1990Q3 to 2318 in 1998Q3. The total number of stock-quarter pairs is 172,936.

Both sets of PIN estimates show a clear downward time trend. The quarterly median EMO estimate is 0.197 in 1983Q1; it gradually goes down to 0.133 in 2004Q4, and reaches the lowest value of 0.123 in 2004Q2. Similarly, the quarterly median LR estimate is 0.187 in 1983Q1; it gradually goes down to 0.138 in 2004Q4, and reaches the lowest value of 0.127 in 2004Q2. We observe the same pattern in the quarterly means.

Fig. 1 presents a notable pattern concerning the difference between the two sets of PIN estimates. In every quarter prior to 1999Q1, the median difference between the two estimates (i.e., the EMO estimate minus the LR estimate of the same stock) is significantly positive at the 1% level. The quarterly median difference ranges from 0.003 to 0.012. However, from 2001Q1 onward, the quarterly median difference is significantly negative at the 1% level in every quarter, ranging from −0.011 to −0.003.

Next, we divide stocks into three groups based on the average daily number of trades for each stock in a quarter. Stocks in the top group are actively traded, whereas stocks in the bottom group are inactively traded. The number of stocks in each group in a quarter ranges from 580 to 773. Fig. 2 plots the cross-sectional median of the daily number of trades in each quarter for the three groups. There is a dramatic increasing trend in the daily number of trades, particularly for the actively traded stocks. At the beginning of our sample period (i.e., 1983Q1), the average daily number of trades is 5.6 for the median inactively traded stock, 16.5 for the median

[5] For more details, please refer to the summary statistics in Table A1 in the online appendix.

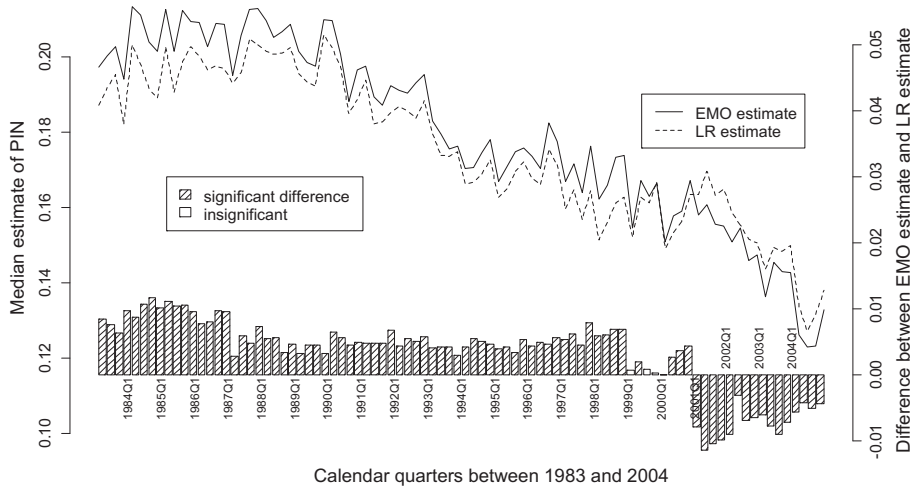

Fig. 1. We obtain two sets of quarterly PIN estimates between 1983 and 2004 using two trade-classification methods. One is the Lee and Ready (LR, 1991) algorithm with a five-second adjustment for delay in reported trade time. The other is the EMO (EMO, 2000) algorithm without time adjustment. The two lines represent the median of the two sets of PIN estimates in each calendar quarter. The bar plots represent the median difference between EMO estimate and LR estimate across stocks. The bars are shaded to represent statistical significance at the 1% level.

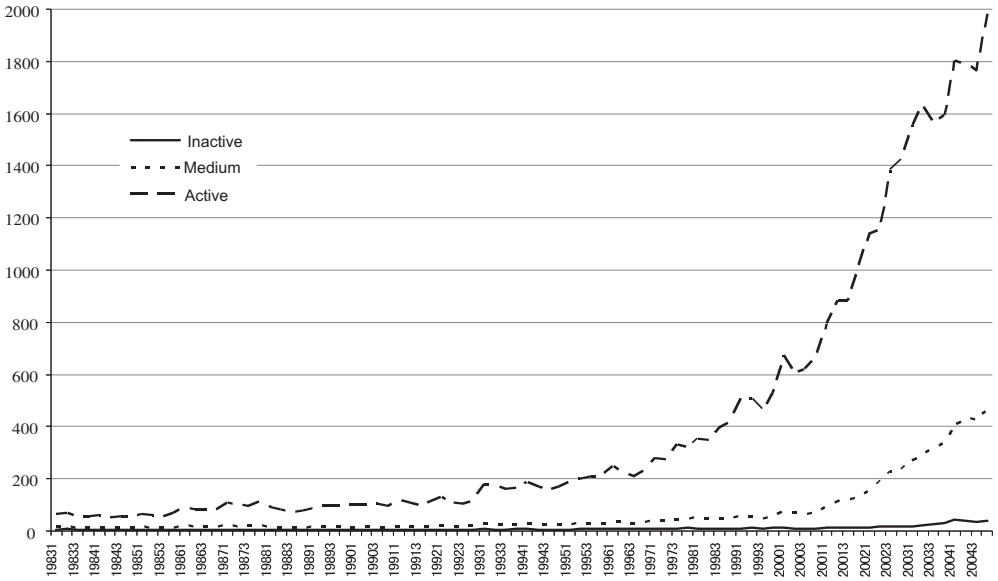

Fig. 2. We sort stocks in our sample by the average daily number of trades in each quarter into three equal-size groups. Actively traded stocks are in the top group, inactively traded stocks are in the bottom group, and the remaining stocks are in the medium group. This figure plots the cross-sectional median of the average daily number of trades for stocks in the three groups in each calendar quarter.

stock in the middle group, and 66.3 for the median actively traded stock. By 2004Q4, the average daily number of trades is 39.3 for the median inactively traded stock, 463.4 for the median stock in the middle group, and 1983.2 for the median actively traded stock.

Fig. 3 plots the cross-sectional average of the EMO estimates and the LR estimates in each quarter for actively and inactively traded stocks. The top panel is for actively traded stocks and the bottom panel for inactively traded stock. The top panel shows that, for actively traded stock, the EMO estimates are, on average, greater than the LR estimates in the quarters prior to 2001, but are smaller in the quarters starting from 2001 onwards. There is no such pattern in the bottom panel for inactively traded stock.

Fig. 3 also illuminates statistical significance of the difference between the two sets of PIN estimates.[6] The shaded bars in Fig. 3 represent a significant difference at the 1% level, whereas the white bars indicate that the difference is not significant at the 1% level. It is evident that, since 2001, the EMO estimates are significantly smaller than the LR estimates for actively traded stock, whereas there is no significant difference between these two sets of estimates for inactively traded stock. The evidence in Fig. 3 suggests that the five-second time adjustment results in an upward bias in the LR estimates for actively traded stock. In the following section, we conduct simulations to investigate whether trade misclassification can result in an overestimation of the PIN measure.

3.2. Simulations

We conduct simulations to examine the impact of trade misclassification on the estimation of PIN.[7] We randomly generate buy and sell trades for 60 days from the Easley et al. (2002) model

[6] The detailed results are reported in Table A2 in the online appendix.

[7] We thank an anonymous reviewer for suggesting this simulation study.

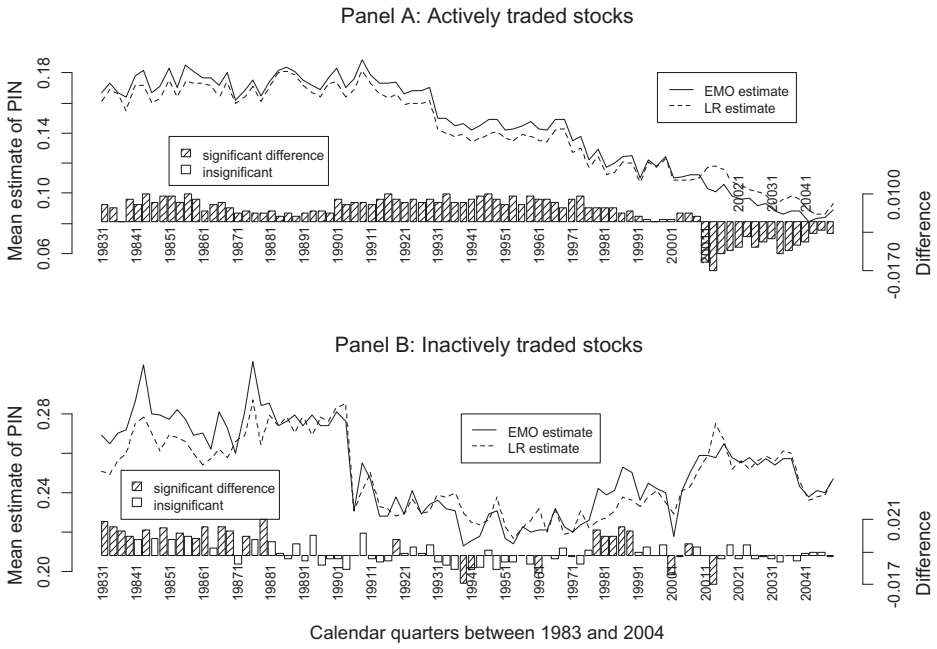

Fig. 3. We sort stocks in our sample by the average daily number of trades in each quarter into three equal-size groups. Actively traded stocks are in the top group, inactively traded stocks are in the bottom group, and the remaining stocks are in the medium group. Panel A is for actively traded stocks, and Panel B is for inactively traded ones. In each panel, the two lines represent the average of the two sets of PIN estimates and the bar plots represent the average difference between EMO estimate and LR estimate. The shaded bars represent statistical significance at 1% level.

Table 1
Simulation results.

Misclassification rate	Misclassification I: Random		Misclassification II: Reducing imbalance		Misclassification III: Enlarging imbalance	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Panel A: True parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.33, 31, 22, 24)$; *True PIN = 0.15874*						
10%	−0.0199	0.0323	−0.0190	0.0319	0.1275	0.1525
15%	−0.0301	0.0384	−0.0290	0.0374	0.3264	0.3403
Panel B: True parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.33, 31, 12, 34)$; *True PIN = 0.15874*						
10%	−0.0188	0.0312	−0.0191	0.0317	0.0494	0.0647
15%	−0.0302	0.0381	−0.0288	0.0372	0.0743	0.0917
Panel C: True parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.31, 21, 11, 13)$; *True PIN = 0.19679*						
10%	−0.0243	0.0391	−0.0254	0.0400	0.1108	0.1311
15%	−0.0386	0.0482	−0.0376	0.0473	0.3459	0.3585
Panel D: True parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.31, 21, 6, 18)$; *True PIN = 0.19679*						
10%	−0.0253	0.0409	−0.0245	0.0392	0.0599	0.0785
15%	−0.0383	0.0484	−0.0383	0.0478	0.0886	0.1113

We run simulations to examine the impact of trade misclassification on the estimation of PIN. This table shows summary statistics based on the simulations. We randomly generate buyer- and seller-initiated trades for 60 days from the Easley et al. (2002) model in Eq. (1). We consider three misclassification schemes under which a fixed proportion of trades (i.e., the misclassification rate) are misclassified. Under the Misclassification I scheme, misclassification occurs randomly. Under the Misclassification II scheme, misclassification occurs such that the daily imbalance between buy and sell trades decreases. Under the Misclassification III scheme, misclassification occurs such that the daily imbalance between buy and sell trades increases. Panels A to D correspond to four sets of true parameter values. We repeat simulations for two misclassification rates: 10% and 15%. We estimate the parameters from the misclassified trades for 60 days using the Lin and Ke (2011) factorization function in Eq. (5) and the Yan and Zhang (2012) algorithm to control for boundary solutions. We run 1000 simulations under each simulation setting and calculate two statistics across the 1000 simulations. The bias represents the average difference between the 1000 simulated PIN estimates and the true PIN value, and the root mean squared error (RMSE) is the square root of the average of the squared differences.

presented in Eq. (1). On any day, trade misclassification will result in one and only one of two consequences: it either reduces or increases the daily imbalance between buy and sell trades. For example, on a particular day, the number of buy trades is 100, the number of sell trades is 50, and thus the buy–sell imbalance is 50. Suppose that 10% of these trades are misclassified. If the number of misclassified buy trades is 10 and the number of misclassified sell trades is five, misclassification causes the number of buy trades and the number of sell trades to be 95 and 55, respectively, and hence reduces the buy–sell imbalance to 40. However, if the number of misclassified buy trades is five and the number of misclassified sell trades is 10, misclassification results in buy trades and sell trades that number 105 and 45, respectively, and hence increases the buy–sell imbalance to 60. The dichotomous effect of misclassification also occurs when there are more sell trades than buy trades.

Our simulations consider three misclassification schemes under which a fixed proportion of trades are misclassified. Under the Misclassification I scheme, misclassification occurs randomly. This implies that the proportion of misclassified trades (i.e., the misclassification rate) may vary between days, that is, misclassification may reduce the daily buy–sell imbalance on some days but enlarge it on others. Under the Misclassification II scheme, the

Table 2
Quarterly variation in PIN.

Calendar quarter # of stock-quarters		1st Quarter 39379	2nd Quarter 41291	3rd Quarter 41139	4th Quarter 40971
ΔPIN	Mean	−0.0019***	0.0011**	0.0010**	−0.0012**
		(−3.81)	(2.21)	(2.00)	(−2.38)
	Median	−0.0049***	0.0004	0.0004	0.0005
		(−1328.50)	(131.50)	(114.50)	(123.00)
$\Delta\alpha$	Mean	0.0036***	0.0002	−0.0004	−0.0005
		(3.21)	(0.15)	(−0.35)	(−0.48)
	Median	0.0011	−0.0035***	0.0000	0.0061***
		(146.00)	(−422.50)	(−44.50)	(719.00)
$\Delta(\varepsilon_b + \varepsilon_s)/\mu$	Mean	0.084***	−0.031***	−0.018***	0.008**
		(25.19)	(−9.59)	(−5.53)	(2.44)
	Median	0.059***	−0.026***	−0.010***	0.018***
		(1806.50)	(−890.50)	(−349.50)	(609.00)

This table reports the mean and median quarterly changes of the probability of informed trading (PIN), the probability of an information event (α), and the ratio of the daily arrival rate of uninformed trades to the daily arrival rate of informed trades ($(\varepsilon_b + \varepsilon_s)/\mu$). The quarterly change of a variable V is $\Delta V_t = (V_t - V_{t-1})$, where V_t is the variable in quarter t. We study the changes in the four calendar quarters, separately. The mean and median are calculated after removing the extreme 1% observations at both tails. The period extends from the second quarter of 1983 (1983Q2) to the fourth quarter of 2004 (2004Q4). The t-statistic and the sign statistic are shown in parentheses.
* Indicate significance at 10%.
** Indicate significance at 5%.
*** Indicate significance at 1%.

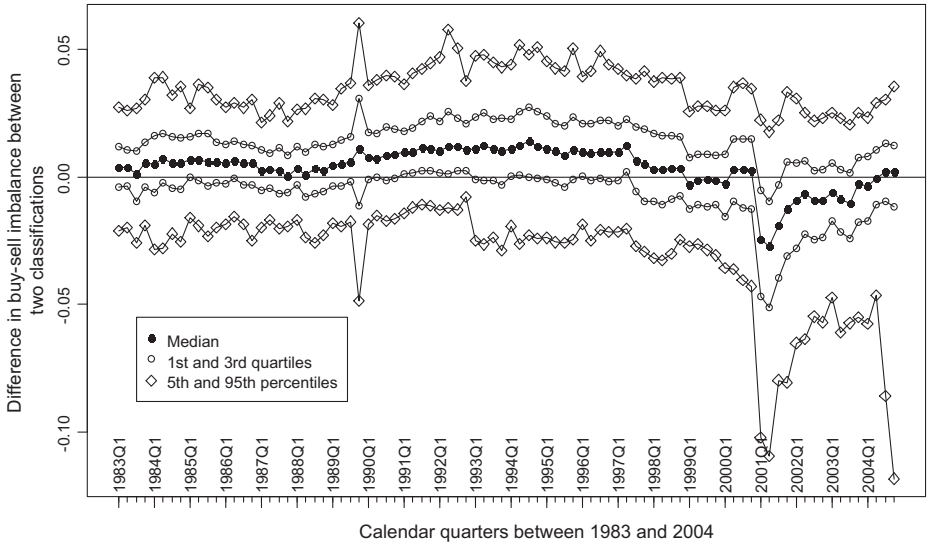

Fig. 4. We sort stocks in our sample by the average daily number of trades in each quarter into three equal-size groups and the stocks in the top group are actively traded. The number of actively traded stocks in a quarter ranges from 580 to 773. On each trading day, for each actively traded stock, we calculate the buy–sell imbalance as the *absolute* difference between the number of buy trades and the number of sell trades divided by the sum of buy and sell trades. We then calculate the difference in the buy–sell imbalance between the EMO without-time-adjustment classification and the Lee-Ready with-five-second-adjustment classification. For each stock, we calculate the average of the daily differences in the buy–sell imbalance over all days in the same calendar quarter. This figure plots the median, the first and third quartiles, and the 5th and 95th percentiles of the distribution of the average difference across all actively traded stocks in each quarter.

misclassification rate is the same each day and misclassification occurs such that the daily buy–sell imbalance is reduced. Under the Misclassification III scheme, the misclassification rate is the same each day and misclassification increases the daily buy–sell imbalance.

We run simulations with four different sets of true values for the five parameters in Eq. (1). The first set includes true parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.33, 31, 22, 24)$, and the true PIN value is 0.15874. We choose these values from the cross-sectional *means* that are reported in Table 2 of Easley et al. (2002). The second set includes true parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.33, 31, 12, 34)$, and the true PIN value is 0.15874. The second set is modified from the first set such that there is a large difference between the daily arrival rate of buy trades ε_b and the daily arrival rate of sell trades ε_s. The third set includes true parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.31, 21, 11, 13)$, and the true PIN value is 0.19679. We choose these values from the cross-sectional *medians* that are shown in Table 2 of Easley et al. (2002). Again, we modified the third set to obtain the fourth set, which includes the true parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s) = (0.28, 0.31, 21, 6, 18)$, and the true PIN value is 0.19679. For the first and third sets of parameters, the arrival rates of buy and sell trades are close, resembling the microstructure model used in Boehmer et al. (2007). For the second and fourth sets of parameters, the arrival rates of buy and sell trades are wide apart, resembling the model used in Easley et al. (2002, 2010), Duarte et al. (2008), and Brown et al. (2004).

We repeat simulations for two misclassification rates 10% and 15%. We estimate the parameters from the misclassified trades for 60 days using the Lin and Ke (2011) factorization function in Eq. (5) and the Yan and Zhang (2012) algorithm to control for boundary solutions. We run 1000 simulations under each simulation setting and calculate two summary statistics across the 1000

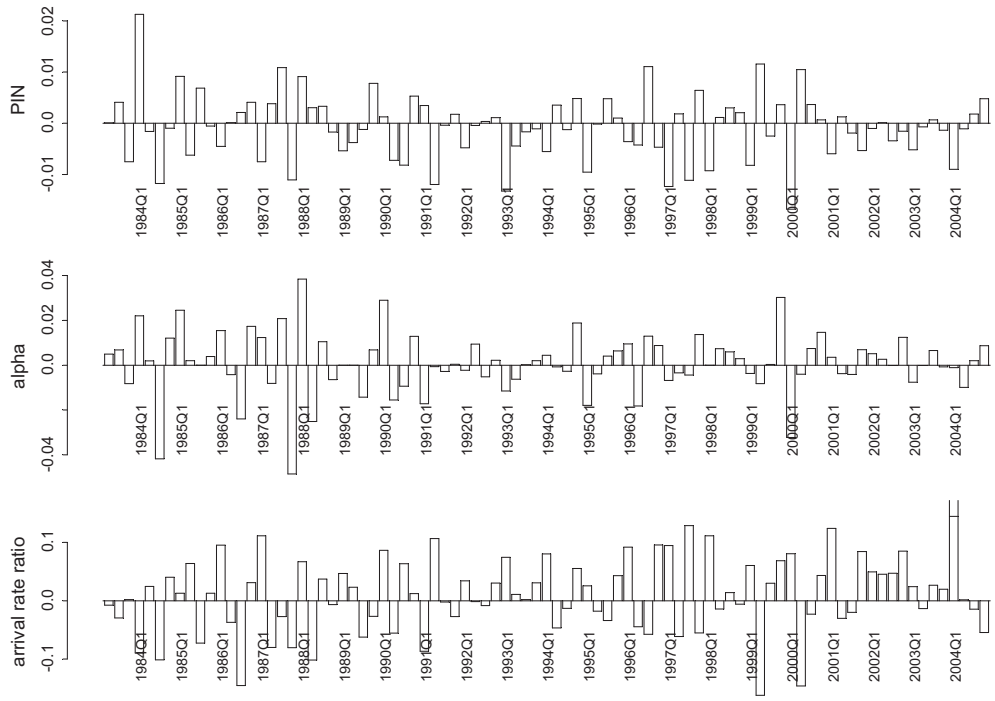

Fig. 5. Quarterly changes in PIN, alpha, and the ratio of the uninformed arrival rate to the informed arrival rate.

Table 3
Year-end tax-loss selling and quarterly change in PIN.

		Fourth Quarter			Subsequent First Quarter		
		Losers	Winners	L minus W	Losers	Winners	L minus W
# of stocks		12,390	17,062		11,812	15,733	
Informed trades $\Delta\mu$ (%)	Mean	40.3	22.7	15.15***	18.9	26.8	−8.18***
	Median	14.3	5.6	12.80***	−1.9	9.0	−15.49***
Uninformed sell$\Delta\varepsilon_s$ (%)	Mean	25.5	12.9	21.56***	0.8	24.7	−40.19***
	Median	15.9	5.5	25.84***	−5.0	17.2	−46.42***
Uninformed buy $\Delta\varepsilon_b$ (%)	Mean	25.8	16.5	11.08***	11.9	29.2	−23.56***
	Median	11.7	6.3	11.45***	−0.3	18.1	−32.55***
Uninformed-to-informed ratio $\Delta(\varepsilon_b+\varepsilon_s)/\mu$ (%)	Mean	12.5	9.5	3.23***	7.9	16.0	−13.80***
	Median	1.5	1.3	0.00	−1.0	8.5	−18.54***
Probability of information events $\Delta\alpha$ (%)	Mean	36.1	23.8	5.27***	28.3	25.4	1.54
	Median	2.3	2.0	0.55	−1.3	1.4	−4.62***
Probability of informed trading ΔPIN (%)	Mean	13.4	8.3	5.18***	10.1	4.1	9.05***
	Median	1.3	0.0	2.24**	−1.0	−5.5	8.99***

This table shows evidence of the relationship between year-end tax-loss selling and quarterly change in PIN. We classify a stock as a winner of the year if its cumulative return from February to November in the year is greater than 10% and as a loser if the return is less than −10%. We study the quarterly percentage changes of six variables: the probability of informed trading (PIN), the probability of an information event (α), the daily arrival rate of informed trades (μ), the daily arrival rates of uninformed buy trades (ε_b), the daily arrival rates of uninformed sell trades (ε_s), and the ratio of the uninformed trades to the informed trades ($(\varepsilon_b+\varepsilon_s)/\mu$). The quarterly percentage change of a variable V is defined as $\Delta V_t = (V_t - V_{t-1})/V_{t-1}$, where V_t is the variable's value in quarter t. We calculate the mean and median quarterly changes of the six variables for winners and losers separately in the fourth quarter and the subsequent first quarter. The mean and median are calculated after removing the extreme 1% observations at both tails. The two-sample t-test and the Wilcoxon rank sum test are used to test the significance of the mean and median difference between losers and winners in the column "L minus W".

simulations. The *bias* is the difference between the average of the 1000 simulated PIN estimates and the true PIN value, and the *root mean squared error* (RMSE) is the square root of the average of the squared differences.

Table 1 reports the bias and RMSE under each simulation setting. A few patterns are evident in Table 1. First, the bias is negative under both Misclassification I and II; however, it is positive under Misclassification III. A negative bias implies that the estimated value is smaller than the true value, whereas a positive bias implies that the estimated value is greater than the true value. Hence, the simulations tell us that there is an underestimation bias when misclassification occurs randomly or when misclassification shrinks the buy–sell imbalance, but there is an overestimation bias when misclassification increases the buy–sell imbalance. This is true for all four sets of parameter settings in Panels A to D.

Second, the magnitude of underestimation is similar under both Misclassification I and II, but the magnitude of overestimation under Misclassification III is significantly larger than the magnitude of underestimation under Misclassification I and II. For example, in Panel A of Table 1, for the misclassification rate 10% and the true PIN 0.15874, the bias under both Misclassification I and II is close to −0.02, but the bias under Misclassification III is almost 0.13.

Table 4
Comparison between Hvidkjaer's PIN estimates and ours.

		Our PIN estimate			HPIN			Ours – HPIN		
Year	N	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
1983	2080	0.208	0.198	0.065	0.224	0.213	0.071	−0.0159***	−0.0139***	0.051
1984	2035	0.221	0.207	0.077	0.209	0.199	0.066	0.0121***	0.0057***	0.064
1985	2011	0.221	0.207	0.074	0.217	0.209	0.065	0.0039***	−0.0031***	0.058
1986	1973	0.216	0.206	0.070	0.219	0.211	0.065	−0.0035***	−0.0055***	0.053
1987	1977	0.218	0.207	0.072	0.222	0.215	0.069	−0.0039***	−0.0051***	0.053
1988	1988	0.227	0.213	0.081	0.221	0.213	0.067	0.0062***	−0.0002	0.064
1989	1895	0.224	0.210	0.080	0.217	0.206	0.070	0.0082***	0.0009	0.064
1990	1802	0.218	0.201	0.082	0.225	0.213	0.073	−0.0067***	−0.0067***	0.068
1991	1859	0.208	0.195	0.079	0.225	0.215	0.075	−0.0166***	−0.0092***	0.063
1992	1975	0.206	0.194	0.075	0.221	0.212	0.076	−0.0150***	−0.0084***	0.063
1993	2134	0.195	0.183	0.077	0.206	0.197	0.068	−0.0104***	−0.0133***	0.060
1994	2199	0.189	0.177	0.074	0.202	0.194	0.066	−0.0138***	−0.0157***	0.057
1995	2237	0.186	0.176	0.071	0.201	0.190	0.071	−0.0156***	−0.0151***	0.057
1996	2288	0.188	0.178	0.067	0.199	0.188	0.073	−0.0113***	−0.0074***	0.053
1997	2403	0.179	0.171	0.067	0.190	0.178	0.076	−0.0105***	−0.0045***	0.048
1998	2476	0.177	0.165	0.071	0.182	0.162	0.083	−0.0044***	0.0010*	0.046
1999	2369	0.180	0.170	0.071	0.184	0.164	0.086	−0.0037***	0.0029***	0.049
2000	2259	0.178	0.167	0.075	0.190	0.168	0.092	−0.0108***	−0.0026***	0.054
2001	2080	0.182	0.165	0.089	0.203	0.180	0.097	−0.0210***	−0.0156***	0.048
All	40040	0.199	0.188	0.077	0.207	0.197	0.076	−0.0072***	−0.0059***	0.057

The HPIN represents the annual estimates obtained from Soren Hvidkjaer for the period 1983–2001. Our PIN estimate is the average of a firm's quarterly estimates in the same year. The quarterly estimates in the years between 1983 and 1998 are based on the Lee and Ready (1991) algorithm with a five-second time adjustment, while the quarterly estimates between 1999 and 2004 are based on the Ellis, Michaely, and O'Hara (EMO, 2000) algorithm without time adjustment.
** Indicate significance at 5% respectively.
* Indicate significance at 10%.
*** Indicate significance at 1%.

Moreover, the magnitude of both underestimation and overestimation increases with the misclassification rate.

The bias and the RMSE under Misclassification III are significantly smaller in Panels B and D than their counterparts in Panels A and C. This implies that the overestimation problem is less severe when the arrival rates of buy and sell trades are wide apart than when the two arrival rates are close.

Boehmer et al. (2007) report an underestimation bias in their simulations when misclassification occurs randomly. We confirm their finding under Misclassification I. We extend their work by considering two additional misclassification schemes and find that when misclassification increases the buy–sell imbalance, an overestimation bias exists.

A relevant question is whether misclassification increases the buy–sell imbalance in real data. Finucane (2000) and Odders-White (2000) provide evidence that trades in actively traded stocks tend to be more frequently misclassified, but there is no evidence in the literature on whether and how misclassification affects buys and sells differently. Fig. 3 shows that the EMO estimates for actively traded stocks are, on average, greater than the LR estimates in the quarters before 2001, but are significantly smaller in the first few quarters since 2001. There is no such pattern for inactively traded stocks. This motivates us to examine actively traded stocks to determine the impact of misclassification on the buy–sell imbalance. We sort stocks in our sample by the average daily number of trades in each quarter into three equally sized groups, and the actively traded stocks are in the top group. The number of actively traded stocks in a quarter ranges from 580 to 773.

On each trading day, for each actively traded stock, we compute the buy–sell imbalance as the *absolute* difference between the number of buy trades and the number of sell trades divided by the sum of buy and sell trades. We then calculate the difference in the buy–sell imbalance between the EMO without-time-adjustment classification and the Lee-Ready with-five-second-adjustment classification. For each stock, we take the average of the daily differences in the buy–sell imbalance over all days in the same calendar quarter. Fig. 4 plots the cross-sectional median, the first and third quartiles, and the 5th and 95th percentiles of the average difference in the buy–sell imbalance across all actively traded stocks in each quarter. It is evident that in the first few quarters since 2001, the median difference in the buy–sell imbalance is significantly below zero. Therefore, over half of the actively

Table 5
Descriptive statistics of the variables in monthly return regressions.

	Mean	Median	SD	P5	P25	P75	P95
Panel A: Descriptive statistics							
Return	0.73	0.21	10.68	−15.70	−5.19	5.92	18.82
QTRPIN	0.188	0.176	0.070	0.099	0.139	0.222	0.328
BETA	1.03	1.01	0.35	0.46	0.78	1.26	1.61
LBM	−0.46	−0.41	0.73	−1.75	−0.87	−0.01	0.67
LSIZE	5.86	5.89	1.98	2.52	4.42	7.31	9.05

	Return	QTRPIN	BETA	LBM	LSIZE
Panel B: Correlation coefficients					
Return	1	0.000	−0.020	0.012	0.015
QTRPIN	−0.024	1	0.148	0.233	−0.661
BETA	−0.035	0.176	1	0.028	−0.313
LBM	0.001	0.232	0.030	1	−0.351
LSIZE	0.047	−0.701	−0.291	−0.363	1

This table shows descriptive statistics of the variables used in the monthly Fama–MacBeth cross-sectional regressions. Return is the monthly excess return. The variable QTRPIN is the average of a firm's quarterly PIN estimates in the past four quarters prior to the month in which the cross-sectional return regression is run. The quarterly PIN estimates in the years between 1983 and 1998 are based on the Lee and Ready (1991) algorithm with a five-second time adjustment, whereas the quarterly estimates between 1999 and 2004 are based on the EMO (2000) algorithm without time adjustment. The variable Return is the monthly excess return. The variable BETA is the portfolio beta calculated from the full period using 100 portfolios. The variable LBM is the logarithm of the ratio of book value of common equity to market value of equity at the end of June each year. The variable LSIZE is the logarithm of the year-end market value of equity in million dollars. Panel A contains the time-series averages of the cross-sectional mean, median, standard deviation (SD), the 5th, 25th, 75th, and 95th percentiles of these variables. Panel B shows correlation coefficients; the Pearson correlation coefficients are in the upper-right triangle, whereas the Spearman rank correlation coefficients are in the lower-left triangle.

Table 6
Monthly stock returns and the QTRPIN.

	April 1983 – March 2005	January 1984 – December 1988	January 1989 – December 1993	January 1994 – December 1998	January 1999 – December 2002
Panel A: Fama–MacBeth cross-sectional regressions, the t-statistic in parenthesis					
Intercept	0.105	−0.240	−0.006	−1.464**	0.759
	(0.26)	(−0.33)	(−0.01)	(−2.02)	(0.55)
QTRPIN	1.414**	2.145**	0.224	3.697***	0.892
	(2.21)	(2.64)	(0.32)	(2.88)	(0.37)
BETA	−0.136	−0.823*	0.125	0.069	−0.399
	(−0.48)	(−1.85)	(0.23)	(0.24)	(−0.36)
LBM	0.203***	0.281*	−0.118	0.306**	0.192
	(2.69)	(1.88)	(−0.70)	(2.46)	(0.86)
LSIZE	0.098*	0.210**	0.074	0.264***	−0.023
	(1.95)	(2.60)	(0.71)	(3.32)	(−0.13)
Panel B: Fama–MacBeth cross-sectional regressions with the Litzenberger and Ramaswamy (1979) weighted least square adjustment, the adjusted *t*-statistic in parenthesis					
Intercept	−0.234	−0.330	−0.265	−1.525**	0.825
	(−0.70)	(−0.52)	(−0.48)	(−2.30)	(0.64)
QTRPIN	2.022***	2.242***	0.724	3.894***	3.105
	(4.08)	(2.85)	(1.17)	(3.63)	(1.38)
BETA	−0.594***	−1.151***	−0.445	0.011	−1.545
	(−2.71)	(−2.79)	(−0.95)	(0.04)	(−1.67)
LBM	0.279***	0.337**	−0.071	0.347***	0.257
	(4.04)	(2.28)	(−0.46)	(2.91)	(1.25)
LSIZE	0.156***	0.221***	0.127	0.258***	0.002
	(3.77)	(2.90)	(1.51)	(3.36)	(0.01)

This table shows the results from the Fama–MacBeth cross-sectional regressions in five time periods. The whole period is from April 1983 to March 2005, and the four sub-periods are 1984–1988, 1989–1993, 1994–1998, and 1999–2002. The variable QTRPIN is the average of a firm's quarterly PIN estimates in the past four quarters prior to the month in which the cross-sectional return regression is run. The quarterly PIN estimates in the years between 1983 and 1998 are based on the Lee and Ready (1991) algorithm with a five-second time adjustment, whereas the quarterly estimates in the years between 1999 and 2004 are based on the EMO (2000) algorithm without time adjustment. The dependent variable Return is the monthly excess return. The variable BETA is the portfolio beta calculated from the full period using 100 portfolios. The variable LSIZE is the logarithm of the year-end market value of equity in million dollars. The variable LBM is the logarithm of the ratio of book value of common equity to market value of equity at the end of June each year.
* Indicate significance at 10%.
** Indicate significance at 5%.
*** Indicate significance at 1%.

traded stocks show a buy–sell imbalance associated with the Lee-Ready with-five-second-adjustment classification significantly larger than the buy–sell imbalance associated with the EMO without-time-adjustment classification. This means that misclassification can increase the buy–sell imbalance for actively traded stocks.

To minimize the impact of the trade misclassification bias, our empirical analysis in the following sections uses the Lee and Ready (1991) method with a five-second time adjustment for the years 1983–1998, and the EMO (2000) method without time adjustment for the years 1999–2004. We choose the break point as 1998/1999 because the differences between the two sets of estimates exhibit a clear paradigm shift from 1998 to 1999 in Fig. 3.

3.3. Quarterly change in PIN

We now examine the change in a firm's PIN from one quarter to the next. Fig. 5 plots the median quarterly change in PIN across firms in every quarter from 1983Q2 to 2004Q4. We observe substantial variation in the PIN measure from one quarter to the next. Additionally, there is a distinct seasonal pattern in the probability of informed trading, that is, the PIN estimates tend to decrease in the first quarter of a year compared with the previous quarter. This is consistent with the findings of Yan and Zhang (2012) during a shorter period 1993–2004.

To further analyze the quarterly change in PIN, we separately examine two key components of PIN. The following transformation of Eq. (3) shows that PIN is positively related to the probability of an information event (i.e., α) and negatively related to the ratio of the daily arrival rate of uninformed trades to the daily arrival rate of informed trades (i.e., $(\varepsilon_b + \varepsilon_s)/\mu$):

$$\mathrm{PIN} = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} = \frac{1}{1 + \frac{1}{\alpha} \cdot \frac{\varepsilon_b + \varepsilon_s}{\mu}}. \quad (7)$$

Table 2 reports separately the statistical tests on quarterly changes in PIN, α, and $(\varepsilon_b + \varepsilon_s)/\mu$ in the four calendar quarters. The mean and median are calculated after removing the extreme 1% observations at both tails. Both the mean and the median quarterly change in PIN are significantly negative in the first quarter.

Both Fig. 5 and Table 2 indicate that the probability of informed trading declines in the first quarter of a year. In an attempt to explain this phenomenon, we hypothesize that the seasonal pattern may be related to tax-loss selling activities at year end.[8] We calculate the cumulative 10-month return from February to November in a year and classify a stock as a winner in the year if its cumulative return is greater than 10% and a loser in the year if the return is less than −10%. We study the quarterly changes of PIN and other parameters in the fourth quarter and the first quarter of the following year. Table 3 shows separately the mean and median of these quarterly changes for winners and losers. The mean and median are calculated after removing the extreme 1% observations at both tails.

We observe significant differences between winners and losers in the fourth quarter and the subsequent first quarter. Specifically,

[8] The year-end tax-loss selling is extensively studied in literature and is found to contribute to the January effect. See, e.g., Grinblatt and Moskowitz (2004), Grinblatt and Keloharju (2003), Poterba and Weisbenner (2001), and references therein.

Table 7
Monthly stock returns and the QTRPIN, excluding January observations.

	April 1983 – March 2005	February 1984 – December 1988	February 1989 – December 1993	February 1994 – December 1998	February 1999 – December 2002
Panel A: Fama–MacBeth cross-sectional regressions, the t-statistic in parenthesis					
Intercept	−0.192	−0.815	−0.502	−1.793**	0.723
	(−0.47)	(−1.13)	(−0.85)	(−2.36)	(0.51)
QTRPIN	1.719**	2.830***	0.504	4.388***	1.076
	(2.54)	(3.53)	(0.74)	(3.26)	(0.41)
BETA	−0.536**	−1.199***	−0.377	−0.059	−1.335
	(−2.13)	(−2.75)	(−0.76)	(−0.19)	(−1.49)
LBM	0.161**	0.228	−0.278*	0.306**	0.184
	(2.10)	(1.57)	(−1.81)	(2.34)	(0.79)
LSIZE	0.172***	0.272***	0.183**	0.305***	0.105
	(3.54)	(3.43)	(2.08)	(3.67)	(0.62)
Panel B: Fama–MacBeth cross-sectional regressions with the Litzenberger and Ramaswamy (1979) weighted least square adjustment, the adjusted t-statistic in parenthesis					
Intercept	−0.474	−0.765	−0.466	−1.812**	0.740
	(−1.37)	(−1.19)	(−0.87)	(−2.61)	(0.56)
QTRPIN	2.282***	2.775***	0.850	4.491***	3.140
	(4.40)	(3.54)	(1.35)	(4.06)	(1.31)
BETA	−0.811***	−1.441***	−0.679	−0.104	−2.071**
	(−3.78)	(−3.58)	(−1.52)	(−0.36)	(−2.47)
LBM	0.263***	0.286*	−0.146	0.350***	0.253
	(3.69)	(2.01)	(−0.94)	(2.80)	(1.18)
LSIZE	0.201***	0.273***	0.176**	0.292***	0.085
	(4.84)	(3.66)	(2.22)	(3.65)	(0.56)

This table reports the results from the Fama–MacBeth cross-sectional regressions in five time periods, excluding January observations. The whole period is from April 1983 to March 2005, and the four sub-periods are 1984–1988, 1989–1993, 1994–1998, and 1999–2002. The variable QTRPIN is the average of a firm's quarterly PIN estimates in the past four quarters prior to the month in which the cross-sectional return regression is run. The quarterly PIN estimates in the years between 1983 and 1998 are based on the Lee and Ready (1991) algorithm with a five-second adjustment, whereas the quarterly estimates in the years between 1999 and 2004 are based on the EMO (2000) algorithm without time adjustment. The dependent variable Return is the monthly excess return. The variable BETA is the portfolio beta calculated from the full period using 100 portfolios. The variable LSIZE is the logarithm of the year-end market value of equity in million dollars. The variable LBM is the logarithm of the ratio of book value of common equity to market value of equity at the end of June each year.
* Indicate significance at 10%.
** Indicate significance at 5%.
*** Indicate significance at 1%.

in the fourth quarter, losers experience a significantly greater increase in both informed and uninformed trading than winners. This is consistent with the year-end tax-loss selling hypothesis, which suggests that losers attract a lot more trading than winners at year end. In the subsequent first quarter, the majority of winners experience a substantial increase in both informed and uninformed trading, whereas more than half of losers experience no increase in trading. This implies that, in the first quarter, investors ignore the previous year's losers and focus on winners instead. As a result, the decrease in the winners' PIN in the first quarter is, on average, substantially greater than that in the losers' PIN.

The above analysis is only a preliminary step toward understanding the forces that drive the inter-temporal change in a firm's information environment captured by the PIN measure. The results in Table 3 suggest that a comprehensive study would be fruitful in this line of inquiry. We intend to pursue such a comprehensive study in the future. We turn to the empirical PIN-return relationship in the next section.

4. Evidence on the PIN-return relationship

Easley et al. (2002) find that monthly returns are significantly and positively related to the PIN measure of information asymmetry in the Fama–MacBeth cross-sectional regressions during the period 1984–1998. Mohanram and Rajgopal (2009) study the PIN-return relationship using Hvidkjaer's PIN estimates for four sub-periods: 1984–1988, 1989–1993, 1994–1998, and 1999–2002, and find that it is significant only in the first sub-period 1984–1988. In the following, we use our PIN estimates to reexamine the PIN-return relationship over a longer period from April 1983 to March 2005.

4.1. Monthly returns and Hvidkjaer's PIN estimates

We compare the average of our quarterly estimates in the same year with the annual estimates from 1983 to 2001 that Soeren Hvidkjaer provided.[9] By averaging quarterly estimates in the same year, we smooth out the seasonal variation documented in Section 3.3. Table 4 reports the cross-sectional mean, median, and standard deviation for each year between 1983 and 2001. Our estimates are, on average, significantly lower than Hvidkjaer's estimates for most years. Note that our PIN estimates are free of the biases associated with trade misclassification (Boehmer et al., 2007), floating-point exception (Lin and Ke, 2011), and boundary solutions (Yan and Zhang, 2012).

We first run the Fama–MacBeth regressions with Hvidkjaer's estimates to provide a benchmark for assessing the Fama–MacBeth regression results based on our own estimates. Because Hvidkjaer's PIN estimates (HPIN) are available at the end of each year from 1983 to 2001, we run the cross-sectional regression for each month between January 1984 and December 2002, inclusively. We use a firm's HPIN at the end of year $t-1$ for the 12 months in year t.

We follow Easley et al. (2002) and Mohanram and Rajgopal (2009) to control for beta, firm size, and book-to-market-equity ratio in the monthly cross-sectional regressions. We calculate betas

[9] We thank Soeren Hvidkjaer for providing his estimates on the website https://sites.google.com/site/hvidkjaer/data.

Table 8
Monthly stock returns and the QTRPIN with controls for the liquidity and momentum effects.

Model	April 1983 – March 2005		January 1984 – December 1988		January 1989 – December 1993		January 1994 – December 1998		January 1999 – December 2002	
	1	2	1	2	1	2	1	2	1	2
Panel A: Fama–MacBeth cross-sectional regressions with the Litzenberger and Ramaswamy (1979) weighted least square adjustment, including January observations										
Intercept	−1.303**	−1.269**	−1.461	−1.759	−1.406	−1.408	−3.098***	−2.459**	0.388	0.350
	(−2.46)	(−2.51)	(−1.29)	(−1.56)	(−1.49)	(−1.67)	(−2.90)	(−2.51)	(0.21)	(0.19)
QTRPIN	1.189**	0.953*	1.835*	1.718*	−0.222	−0.457	2.640**	2.272*	2.494	2.101
	(2.16)	(1.83)	(1.86)	(1.90)	(−0.24)	(−0.53)	(2.15)	(1.84)	(1.17)	(1.05)
BETA	−0.577***	−0.473**	−1.125***	−0.964**	−0.370	−0.194	0.059	−0.008	−1.608*	−1.327*
	(−2.67)	(−2.39)	(−2.76)	(−2.61)	(−0.79)	(−0.41)	(0.22)	(−0.03)	(−1.76)	(−1.75)
LBM	0.286***	0.299***	0.320**	0.331**	−0.075	−0.024	0.362***	0.377***	0.260	0.268
	(4.02)	(4.35)	(2.09)	(2.15)	(−0.46)	(−0.16)	(2.95)	(3.12)	(1.24)	(1.32)
LSIZE	0.340***	0.314***	0.432**	0.443***	0.309**	0.277**	0.511***	0.417***	0.084	0.049
	(4.59)	(4.58)	(2.54)	(2.69)	(2.18)	(2.22)	(3.55)	(3.18)	(0.36)	(0.23)
ILLIQ	0.153***	0.146***	0.166	0.183	0.127	0.117	0.180*	0.129	0.077	0.045
	(2.76)	(2.92)	(1.23)	(1.46)	(1.17)	(1.26)	(1.81)	(1.41)	(0.44)	(0.29)
RET1		−4.532***		−5.648***		−3.150***		−4.283***		−5.354***
		(−9.61)		(−5.72)		(−3.28)		(−4.91)		(−3.99)
RET2to13		0.882***		1.002**		1.467***		1.170***		1.170**
		(5.22)		(2.56)		(4.57)		(3.98)		(2.15)
	April 1983 – March 2005		February 1984 – December 1988		February 1989 – December 1993		February 1994 – December 1998		February 1999 – December 2002	
	1	2	1	2	1	2	1	2	1	2
Panel B: Fama–MacBeth cross-sectional regressions with the Litzenberger and Ramaswamy (1979) weighted least square adjustment, excluding January observations										
Intercept	−1.308**	−1.360**	−1.519	−2.009*	−1.126	−1.237	−3.515***	−2.827***	0.713	0.296
	(−2.36)	(−2.55)	(−1.26)	(−1.68)	(−1.22)	(−1.45)	(−3.14)	(−2.75)	(0.37)	(0.16)
QTRPIN	1.781***	1.411***	2.722***	2.305**	0.087	−0.247	3.403***	2.950**	3.802*	3.328*
	(3.28)	(2.67)	(3.01)	(2.56)	(0.10)	(−0.28)	(2.80)	(2.36)	(1.87)	(1.71)
BETA	−0.798***	−0.668***	−1.418***	−1.205***	−0.623	−0.445	−0.055	−0.117	−2.132**	−1.764**
	(−3.76)	(−3.37)	(−3.56)	(−3.27)	(−1.41)	(−1.01)	(−0.19)	(−0.41)	(−2.55)	(−2.37)
LBM	0.263***	0.275***	0.257*	0.255*	−0.168	−0.101	0.367***	0.372***	0.249	0.261
	(3.58)	(3.89)	(1.71)	(1.74)	(−1.05)	(−0.66)	(2.86)	(2.93)	(1.13)	(1.23)
LSIZE	0.339***	0.323***	0.397**	0.442**	0.277*	0.260**	0.561***	0.456***	0.081	0.082
	(4.39)	(4.47)	(2.27)	(2.57)	(1.94)	(2.02)	(3.70)	(3.29)	(0.32)	(0.36)
ILLIQ	0.111**	0.119**	0.091	0.140	0.061	0.072	0.192*	0.138	−0.015	−0.003
	(2.01)	(2.28)	(0.68)	(1.07)	(0.61)	(0.79)	(1.80)	(1.40)	(−0.09)	(−0.02)
RET1		−3.776***		−4.888***		−2.341***		−3.456***		−4.552***
		(−8.28)		(−5.07)		(−2.71)		(−4.11)		(−3.51)
RET2to13		0.974***		1.249***		1.576***		1.282***		1.264***
		(5.67)		(3.39)		(4.77)		(4.11)		(2.32)

This table shows the results from the Fama–MacBeth cross-sectional regressions with controls for the liquidity and momentum effects. January observations are included in Panel A, but excluded in Panel B. The whole period is from April 1983 to December 2004, and the four sub-periods are 1984–1988, 1989–1993, 1994–1998, and 1999–2002. The variable QTRPIN is the average of a firm's quarterly PIN estimates in the past four quarters prior to the month in which the cross-sectional return regression is run. The quarterly PIN estimates in the years between 1983 and 1998 are based on the Lee and Ready (1991) algorithm with a five-second time adjustment, whereas the quarterly estimates in the years between 1999 and 2004 are based on the EMO (2000) algorithm without time adjustment. The dependent variable Return is the monthly excess return. The variable BETA is the portfolio beta calculated from the full period using 100 portfolios. The variable LSIZE is the logarithm of the year-end market value of equity in million dollars. The variable LBM is the logarithm of the ratio of book value of common equity to market value of equity at the end of June each year. The variable RET1 represents the return in month $t-1$ and RET2to13 represents the return over the previous two to thirteen months. The variable ILLIQ is the logarithm of Amihud's liquidity measure. The adjusted t-statistic based on the Litzenberger and Ramaswamy (1979) weighted least square method is in parentheses.

* Indicate significance at 10%.
** Indicate significance at 5%.
*** Indicate significance at 1%.

according to Fama and French (1992). Pre-ranking portfolio betas are estimated for individual stocks using monthly returns in at least two to five years. We regress these stock returns on the contemporaneous and lagged value-weighted CRSP NYSE/AMEX index to obtain two coefficients. Pre-ranking portfolio beta is the sum of the two coefficients used to adjust for nonsynchronous trading. We then sort the stocks into 100 portfolios based on these betas at the end of each year and calculate monthly equal-weighted portfolio returns. Post-ranking portfolio betas are estimated using the full period. Portfolio returns are regressed on the contemporaneous and lagged value-weighted CRSP index returns, and again the portfolio beta is the sum of the two coefficients. The variable BETA of any given stock in the 12 months in year t is equal to the beta of the portfolio to which it belongs at the end of year $t-1$.

We follow Daniel and Titman (2006) to calculate the book value of equity from the annual COMPUSTAT data set. The observations with negative book value are removed. We measure the book-to-market-equity ratio in June of year t as the last fiscal year's book value divided by the market capitalization at the end of June and use it for the 12 months from July of year t until June of year $t+1$. The variable LBM is the natural logarithm of the book-to-market-equity ratio. We measure firm size using the market capitalization in millions of US dollars at the end of year $t-1$ and use it for the 12 months in year t. The variable LSIZE is the natural logarithm of firm size. We winsorize each independent variable at the first and 99th percentiles in each month.

We run the Fama–MacBeth regressions in five time periods: the whole period 1984–2002 and the four sub-periods 1984–1988, 1989–1993, 1994–1998, and 1999–2002.[10] For the whole period 1984–2002, the coefficient of HPIN is positive and significant, which is consistent with the original finding of Easley et al. (2002). The sign and magnitude of the other variable coefficients are also similar to those of Easley et al. (2002). Moreover, we confirm Mohanram and Rajgopal (2009)'s finding that the PIN-return relationship is significant only in the first sub-period 1984–1988.

4.2. Monthly returns and our PIN estimates

Next, we re-examine the PIN-return relationship with our quarterly PIN estimates. To control for the misclassification bias documented in Section 3.1, we obtain the quarterly estimates between 1983 and 1998 based on the Lee and Ready (1991) algorithm with a five-second time adjustment and the quarterly estimates between 1999 and 2004 based on the EMO (2000) algorithm without time adjustment. To smooth out the seasonal variation observed in Table 2, we use the average of quarterly PIN estimates for the four quarters prior to the month in which the cross-sectional return regression is run.

We calculate, in each month, the cross-sectional mean, median, standard deviation, the 5th, 25th, 75th, and 95th percentiles of each variable, and the correlation coefficients between these variables. Panels A and B of Table 5 report the time-series average of these descriptive statistics. The variable QTRPIN represents the average of quarterly PIN estimates. We observe that QTRPIN has a significant positive correlation with the variables BETA and LBM, but a substantially negative correlation with the variable LSIZE. The variable LSIZE has a substantially negative correlation with the variables BETA and LBM.

Table 6 reports the results of the Fama–MacBeth regressions in five time periods: the whole period from April 1983 to March 2005 and the four sub-periods 1984–1988, 1989–1993, 1994–1998, and 1999–2002. Easley et al. (2002) find a stronger PIN-return relationship using the Litzenberger and Ramaswamy (1979) weighted least square method to adjust for time-varying precision in monthly regression estimates. Panel A of Table 6 shows the results using the conventional t-statistic, whereas Panel B reports the results of the Fama–MacBeth regressions with the Litzenberger and Ramaswamy (1979) adjustment.

For the whole period from April 1983 to March 2005, the coefficient of the variable QTRPIN is significantly positive, whereas the coefficients of the other independent variables are similar to those of Easley et al. (2002). Notably, Table 6 finds that QTRPIN is significantly positive for the sub-period 1984–1988 and the sub-period 1994–1998. This is in contrast to Mohanram and Rajgopal (2009) who find that the PIN-return relationship is significant only in the first sub-period 1984–1988. Although PIN is insignificant for the sub-period 1999–2002, it is noteworthy that none of the independent variables, including firm size and book-to-market-equity ratio, is significant in this sub-period.

4.3. The robustness of the PIN-return relationship

Kang (2011) finds a January effect in the PIN-return relationship. He presents evidence that stock returns decrease with PIN in January, but increase with PIN in other months. Table 7 shows the results from the Fama–MacBeth regression analysis after excluding January observations. The results in Panel B are adjusted for time-varying precision in monthly regression estimates according to Litzenberger and Ramaswamy (1979). We make two observations in Table 7. First, consistent with that in Table 6, the variable QTRPIN has a significantly positive coefficient for not only the whole period but also the two sub-periods 1984–1988 and 1994–1998. This provides additional evidence in support of the positive PIN-return relationship. Second, the positive coefficient of the variable QTRPIN is larger in Table 7 than that in Table 6. This is consistent with Kang (2011)'s finding that PIN and returns are negatively related in January, but positively related in other months.

The empirical asset pricing literature has discovered several variables commonly used to explain asset returns. They include beta, size, book-to-market-equity ratio (Fama and French, 1993), momentum (Jegadeesh and Titman, 1993; Carhart, 1997; Grundy and Martin, 2001), and liquidity (Amihud, 2002). We use the past one-month return (RET1) and the cumulative return over the previous two to thirteen months (RET2to13) to capture short-term return reversals (Jegadeesh, 1990) and intermediate-term momentum (Jegadeesh and Titman, 1993). We follow Amihud (2002) to calculate the illiquidity measure. Table 8 reports the results from the cross-sectional regressions with these control variables; Panel A is for all months including January and Panel B is for non-January months. The results in both panels are adjusted for time-varying precision in monthly regression estimates according to Litzenberger and Ramaswamy (1979).

We find that the variable QTRPIN continues to have a positive and significant coefficient after we include these control variables. The sub-period results in both Panels A and B show that the coefficient of QTRPIN is significant in the two sub-periods, 1984–1988 and 1994–1998. Moreover, after January observations are excluded, the coefficient of QTRPIN in Panel B is also significant at the 10% level for one additional sub-period 1999–2002.

5. Summary and conclusion

Several recent studies point out that the estimates of the PIN measure of information asymmetry may contain certain biases, because of computer floating-point exception, boundary solutions, and trade misclassification. Such biases may have an impact on

[10] For the sake of brevity, we omit the tables that show these results. They are available upon request.

the positive relationship between PIN and stock returns discovered by Easley et al. (2002). We apply several methods to overcome these estimation biases and improve the quality of PIN estimates. Specifically, we use the Lin and Ke (2011) factorization method to reduce the impact of floating-point exception, and the Yan and Zhang (2012) algorithm to avoid boundary solutions. We evaluate two trade classification methods: the Lee and Ready (1991) algorithm with a five-second adjustment for the delay in reported trade time and the EMO (2000) algorithm without time adjustment. Our empirical and simulation results suggest that we use the first method for the period 1983–1998 and the second method for the period 1999–2004.

We estimate PIN for more than 170,000 stock-quarters and re-examine the PIN-return relationship with the Fama–MacBeth regression methodology over 276 months between April 1983 and March 2005, inclusively. We find stronger evidence for a positive PIN-return relationship than documented by the prior literature. This finding is robust after we control for the January, liquidity, and momentum effects and use the Litzenberger and Ramaswamy (1979) method to adjust for time-varying precision in monthly regression estimates.

The number of studies that use PIN as a measure of information asymmetry is increasing. This study demonstrates that the quality of PIN estimates has an influence on the observed evidence of the PIN-return relationship. In conclusion, we recommend that researchers apply the methods used in this study to obtain high-quality estimates of PIN, which would make their empirical evidence more convincing.

Acknowledgments

The authors would like to thank an anonymous reviewer for valuable comments that help us improve the paper significantly. We also thank Hung Wan Kot, Ji-Chai Lin, Lin Peng, Bohui Zhang, and seminar participants at the Hong Kong Polytechnic University, the 2012 Asian Finance Association Conference, the 2012 China International Conference in Finance, and the 2012 Financial Management Association Conference for helpful comments. Part of the research was done while Yan was affiliated with the Wharton School of the University of Pennsylvania and Zhang was with the Nanyang Business School of the Nanyang Technological University. Zhang gratefully acknowledges financial support from the School of Accounting and Finance in the Hong Kong Polytechnic University (Grant #A-PJ99) and the Hong Kong Government Theme-based Research Project "Enhancing Hong Kong's Future as a Leading International Financial Center". All remaining errors are our own.

References

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5 (1), 31–56.
Bessembinder, H., 2003. Issues in assessing trade execution costs. Journal of Financial Markets 6 (3), 233–257.
Boehmer, E., Grammig, J., Theissen, E., 2007. Estimating the probability of informed trading: does trade misclassification matter? Journal of Financial Markets 10 (1), 26–47.
Brockman, P., Yan, X.S., 2009. Block ownership and firm-specific information. Journal of Banking and Finance 33 (2), 308–316.
Brown, S., Hillegeist, S., Lo, K., 2004. Conference calls and information asymmetry. Journal of Accounting and Economics 37 (3), 343–366.
Carhart, M., 1997. On persistence in mutual fund performance. Journal of Finance 52, 57–82.
Chen, Q., Goldstein, I., Jiang, W., 2007. Price informativeness and investment sensitivity to stock price. Review of Financial Studies 20 (3), 619–650.
Chen, Y., Zhao, H., 2012. Informed trading, information uncertainty, and price momentum. Journal of Banking and Finance 36 (7), 2095–2109.
Daniel, K., Titman, S., 2006. Market reactions to tangible and intangible information. Journal of Finance 61 (4), 1605–1643.
Duarte, J., Han, X., Harford, J., Young, L., 2008. Information asymmetry, information dissemination and the effect of regulation FD on the cost of capital. Journal of Financial Economics 87 (1), 24–44.
Duarte, J., Young, L., 2009. Why is PIN priced? Journal of Financial Economics 91 (2), 119–138.
Easley, D., Hvidkjaer, S., O'Hara, M., 2002. Is information risk a determinant of asset returns? Journal of Finance 57 (5), 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., O'Hara, M., 1997. One day in the life of a very common stock. Review of Financial Studies 10 (3), 805–835.
Easley, D., O'Hara, M., 1992. Time and the process of security price adjustment. Journal of Finance 47 (2), 577–604.
Easley, D., O'Hara, M., 2004. Information and the cost of capital. Journal of Finance 59 (4), 1553–1583.
Ellis, K., Michaely, R., O'Hara, M., 2000. The accuracy of trade classification rules: evidence from Nasdaq. Journal of Financial and Quantitative Analysis 35 (4), 529–551.
Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. Journal of Finance 47 (2), 427–465.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33 (1), 356–372.
Fama, E.F., MacBeth, J.D., 1973. Risk, return and equilibrium: empirical tests. Journal of Political Economy 81, 607–636.
Finucane, T.J., 2000. A direct test of methods for inferring trade direction from intraday data. Journal of Financial and Quantitative Analysis 35 (4), 553–576.
Francis, J., LaFond, R., Olsson, P., Schipper, K., 2004. Cost of equity and earnings attributes. Accounting Review 79 (4), 967–1010.
Francis, J., LaFond, R., Olsson, P., Schipper, K., 2005. The market pricing of accruals quality. Journal of Accounting and Economics 39 (2), 295–327.
Grinblatt, M., Moskowitz, T.J., 2004. Predicting stock price movements from past returns: the role of consistency and tax-loss selling. Journal of Financial Economics 71 (3), 541–579.
Grinblatt, M., Keloharju, M.K., 2003. Tax-loss trading and wash sales. Journal of Financial Economics 71, 51–76.
Grundy, B., Martin, S., 2001. Understanding the nature and the risks and the sources of the rewards to momentum investing. Review of Financial Studies 14 (1), 29–78.
Henker, T., Wang, J.X., 2006. On the importance of timing specifications in market microstructure research. Journal of Financial Markets 9 (2), 162–179.
Hughes, J.S., Liu, J., Liu, J., 2007. Information asymmetry, diversification, and cost of capital. Accounting Review 82 (3), 705–729.
Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. Journal of Finance 45 (3), 881–898.
Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. Journal of Finance 48 (1), 65–91.
Kang, M., 2011. Probability of information-based trading and the January effect. Journal of Banking and Finance 34, 2985–2994.
Lee, C., Ready, M., 1991. Inferring trade direction from intraday data. Journal of Finance 46, 733–746.
Lin, H.W., Ke, W.C., 2011. A computing bias in estimating the probability of informed trading. Journal of Financial Markets 14 (4), 625–640.
Litzenberger, R., Ramaswamy, K., 1979. The effect of personal taxes and dividends on capital asset prices: theory and empirical evidence. Journal of Financial Economics 7 (2), 163–196.
Mohanram, P., Rajgopal, S., 2009. Is PIN priced risk? Journal of Accounting and Economics 47 (3), 226–243.
Odders-White, E.R., 2000. On the occurrence and consequences of inaccurate trade classification. Journal of Financial Markets 3 (3), 259–286.
Odders-White, E.R., Ready, M., 2006. Credit ratings and stock liquidity. Review of Financial Studies 19 (1), 119–157.
Peterson, M., Sirri, E., 2003. Evaluation of the biases in execution cost estimation using trade and quote data. Journal of Financial Markets 6 (3), 259–280.
Poterba, J.M., Weisbenner, S.J., 2001. Capital gains tax rules, tax-loss trading, and turn-of-the-year returns. Journal of Finance 56 (1), 353–368.
Yan, Y., Zhang, S., 2012. An improved estimation method and empirical properties of the probability of informed trading. Journal of Banking and Finance 36 (2), 454–467.