The Sensitivity to Trade Classification Algorithms for Estimating the Probability of Informed Trading

Wen-Chyan Ke

***Abstract*—This study examines the impact of trade classification algorithms on estimating the probability of informed trading (PIN). This study finds that the algorithms themselves may not substantially influence the PIN estimates but the poor performances of these algorithms may have the great impact on the PIN estimates. Moreover, the new proposed adjustment, Q-Method, seems to mitigate the bias caused by the trade misclassification. In addition, the pattern of its estimates responds to the important economic events. With the estimated misclassification rate from Q-Method, this study also finds that the performances of these algorithms are getting poor in recent years.**

***Index Terms*—Informed trading, market microstructure, trade misclassification.**

I. Introduction

This study proposes a remedy, called Q-Method, for estimating the probability of informed trading (hereafter also referred to as PIN) when the misclassification of trades occurs (and results in biased PIN estimates). Easley *et al*. in [1] and [2] develop the PIN measure from their microstructure model, denoted as "PIN Model" hereafter. The Q-Method can mitigate estimation biases for both the PIN Model and its extensions. Moreover, the PIN has been widely employed in securities market studies (e.g., [3]-[7]).

This study focuses on examining the sensitivity of PIN estimation to the trade classification algorithms. In the literature [8] and [9], other algorithms are proposed with the different sorting strategies. The past studies [8] and [9] also have demonstrated that these algorithms have different classification accuracy rates, and significantly result in biased estimates of the effective spreads and the price impacts.

Boehmer *et al*. in [10] show analytically that the inaccurate trade classification leads to downward-biased PIN estimates and that the magnitude of the bias is related to a security's trading intensity. Therefore, the different classification algorithms clearly result in the varied estimation biases of PIN due to their inconsistent accurate rates. Therefore, this study further examines the effectiveness of Q-Method using different algorithms.

The estimation of PIN relies on the intraday information regarding the trade direction. However, the availability of intraday trade and quotes data does not provide information on the trade direction. Therefore, the trade classification algorithms are used to distinguish between buyer- and seller-initiated trades (or *buys* and *sells*). Such information regarding the trade direction is also an important element in determining the price impact of large trade, the effective spread, and many other related questions (e.g., [11], [12]). Namely, the trade classification algorithms have been extensively used in microstructure studies.

For example, there are three algorithms proposed by Lee and Ready (LR) in [13], Eillis *et al.* (EMO) in [9] and Chakrabarty *et al.* (CBNV) in [8], respectively. These algorithms infer the trade classification from trade and quote data by comparing the trade price to previous trade prices or to prevailing quotes. Using the TORQ (Trades, Oreder, Reports, and Quotes) database from NYSE, Lee and Radhakrishna in [14] conduct an evaluation of LR and find an accurate rate of 93% in their select sample. By contrast, using different selection rules, Odders-White in [15] documents an accurate rate of 85% for LR in a different sample of TORQ. With the data of NASDAQ, Eillis *et al.* in [9] find that LR and their proposed EMO algorithms correctly classify 81.95% and 83.74% of the trades, respectively. Chakrabarty *et al.* in [8] refine the EMO algorithm and find that the overall accurate rates of 74.42%, 75.80% and 76.52% for LR, EMO and their proposed CBNV algorithms, respectively, in a sample of NASDAQ stocks that trade on the ECN (electronic communications network). Furthermore, the prior studies [8] and [9] have documented that LR, EMO and CBNV methods result in statistically significant difference in the estimates of actual effective spreads and price impacts.

Adding to literature, this study estimates the PINs with the process of Easley *et al*. in [1] and [2] (hereafter referred to as E-Method) for the three classification algorithms, respectively, and compare the difference in these PIN estimates. According to past studies [8], [9] and [10], this study estimates PIN with E-Method for the LR, EMO and CBNV. Then, this study finds that the three algorithms themselves may not lead to the substantially different PIN estimates but their misclassification rates may. Moreover, this study finds that Q-Method improves the quality of PIN estimates for the three algorithms.

The remaining sections are organizes as the following. First, the relationship between Q- and E-Methods is described. Second, the details of the LR, OME and CBNV algorithms are provided. Next, the empirical procedures are constructed with the TAQ data. Finally, this study provides the discussion and conclusion.

Manuscript received May 25, 2014; revised July 21, 2014. This work was supported in part by the National Science Council of Taiwan for support under Grant NSC101-2410-H-305-034-.

W.-C. Ke is with the Department of Finance and Cooperative Management National Taipei University, No. 151, University Rd., San-Shia District, New Taipei City 23741, Taiwan (e-mail: wenchyan@gm.ntpu.edu.tw).

DOI: 10.7763/IJTEF.2014.V5.404

II. Estimating the Probability of Informed Trading with Q-Method

This study develops the Q-Method, which aims at both PIN Model and extended PIN Models such as [16] and [17]. The PIN estimation procedure proposed by Easley *et al.* in [1] and [2] (the E-Method) does not consider the misclassification of trades and is a special case of Q-Method.

The Q- and E-Methods use different ways to estimate PIN. E-Method uses the numbers of *buys* and *sells* within a day to guess the arrival rates, which does not handle the misclassification. The Q-Method, in contrast, uses the numbers of *buys* and *sells* to guess the adjusted arrival rates, which are consistent with those of imperfectly classified *buys* and *sells*. Moreover, Q-Method adds to the literature with its effectiveness for most extensions of PIN Model. The following provides a snapshot of Q-Method.

To estimate parameters such as PIN, empiricists typically need to adopt the proxies of buy and sell orders. Yet daily proxies, called *buys* and *sells*, are generally unobservable, and may have to be inferred from algorithms that inevitably accompany misclassification (e.g., [8], [9], [13], [15]).

Let B_i and S_i denote the numbers of imperfectly classified *buys* and *sells* on day i, and $\widetilde{B_i}$ and $\widetilde{S_i}$ the numbers of actual and unobserved *buys* and *sells*, for convenience. The notation $\text{Pois}(x|\lambda) \equiv e^{-\lambda}\lambda^x/x!$ denotes the probability density function of Poisson variable x with arrival rate λ.

In the PIN Model, the joint probability density function of Poisson variables $\widetilde{B_i}$ and $\widetilde{S_i}$ could be specified as follows:

$$\begin{aligned} f(\widetilde{B_i}, \widetilde{S_i}|\boldsymbol{\theta}) &= \alpha\delta f(\widetilde{B_i}, \widetilde{S_i}|\boldsymbol{\theta}, \text{bad news}) \\ &+ \alpha(1-\delta) f(\widetilde{B_i}, \widetilde{S_i}|\boldsymbol{\theta}, \text{good news}) \\ &+ (1-\alpha) f(\widetilde{B_i}, \widetilde{S_i}|\boldsymbol{\theta}, \text{no news}) \\ &= \alpha\delta \text{Pois}(\widetilde{B_i}|\varepsilon_b)\text{Pois}(\widetilde{S_i}|\varepsilon_s + \mu) \\ &+ \alpha(1-\delta)\text{Pois}(\widetilde{B_i}|\varepsilon_b + \mu)\text{Pois}(\widetilde{S_i}|\varepsilon_s) \\ &+ (1-\alpha)\text{Pois}(\widetilde{B_i}|\varepsilon_b)\text{Pois}(\widetilde{S_i}|\varepsilon_s), \end{aligned} \tag{1}$$

where α is the probability of an information event; δ and $(1-\delta)$ are the conditional probabilities of bad and good news types, respectively; ε_b and ε_s are the arrival rate of uninformed *buys* and that of uninformed *sells*, respectively; μ is the arrival rate of informed trades; and vector $\boldsymbol{\theta} = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s)$ represents the structural parameters.

When misclassification is present, given a news type on day i (whether bad, good, or no news), the actual arrival rates of the observed Poisson variables B_i and S_i may be derived from those of unobserved $\widetilde{B_i}$ and $\widetilde{S_i}$. For example, if each trade is independently and incorrectly classified with the probability $(1-q)$ (also the misclassification rate), then the number of correct classifications in $\widetilde{B_i}$ *buys* is $B_i^C \approx q\widetilde{B_i} \le \widetilde{B_i}$, and the number of incorrect classifications in $\widetilde{S_i}$ *sells* is $\widetilde{S_i} - S_i^C \approx (1-q)\widetilde{S_i} \ge 0$, where S_i^C represents the number of correctly classified *sells*. Accordingly, $B_i = B_i^C + (\widetilde{S_i} - S_i^C) \approx q\widetilde{B_i} + (1-q)\widetilde{S_i}$.

Given the news type, the arrival rate of imperfectly classified *buys*—the conditional mean of B_i—should be the weighted average arrival rate where the weight q corresponds to the arrival rate of actual *buys*—the conditional mean of $\widetilde{B_i}$—and $(1-q)$ corresponds the arrival rate of actual *sells*—the conditional mean of $\widetilde{S_i}$.

For instance, given $\boldsymbol{\theta}_Q \equiv (\boldsymbol{\theta}, q) = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s, q)$ and that bad news emerges on day i,

$$\begin{aligned} E(B_i|\text{bad news}) &= q\varepsilon_b + (1-q)(\varepsilon_s + \mu) \text{ and} \\ E(S_i|\text{bad news}) &= q(\varepsilon_s + \mu) + (1-q)\varepsilon_b. \end{aligned} \tag{2}$$

Therefore, the following conditional probability density function is obtained:

$$\begin{aligned} f(B_i, S_i|\boldsymbol{\theta}_Q, \text{bad news}) &= \text{Pois}(B_i|q\varepsilon_b + (1-q)(\varepsilon_s + \mu)) \\ &\times \text{Pois}(S_i|q(\varepsilon_s + \mu) + (1-q)\varepsilon_b). \end{aligned} \tag{3}$$

Accordingly, the joint probability density function of B_i and S_i is

$$\begin{aligned} f(B_i, S_i|\boldsymbol{\theta}_Q) &= \alpha\delta f(B_i, S_i|\boldsymbol{\theta}_Q, \text{bad news}) \\ &+ \alpha(1-\delta) f(B_i, S_i|\boldsymbol{\theta}_Q, \text{good news}) \\ &+ (1-\alpha) f(B_i, S_i|\boldsymbol{\theta}_Q, \text{no news}) \\ &= \alpha\delta \text{Pois}(B_i|q\varepsilon_b + (1-q)(\varepsilon_s + \mu)) \\ &\times \text{Pois}(S_i|q(\varepsilon_s + \mu) + (1-q)\varepsilon_b) \\ &+ \alpha(1-\delta)\text{Pois}(B_i|q(\varepsilon_b + \mu) + (1-q)\varepsilon_s) \\ &\times \text{Pois}(S_i|q\varepsilon_s + (1-q)(\varepsilon_b + \mu)) \\ &+ (1-\alpha)\text{Pois}(B_i|q\varepsilon_b + (1-q)\varepsilon_s) \\ &\times \text{Pois}(S_i|q\varepsilon_s + (1-q)\varepsilon_b). \end{aligned} \tag{4}$$

Moreover, with the assumption that the arrivals on each trading day are independent of one another in [1] and [2], the joint log-likelihood of observing a series of (B_i, S_i) over the past I trading days is the sum of the daily log-likelihoods $L(\boldsymbol{\theta}_Q|B_i, S_i) \equiv \log(f(B_i, S_i|\boldsymbol{\theta}_Q))$,

$$L(\boldsymbol{\theta}_Q|\mathbf{D}) \equiv \sum_{i=1}^{I} L(\boldsymbol{\theta}_Q|B_i, S_i) = \sum_{i=1}^{I} \log\left(f(B_i, S_i|\boldsymbol{\theta}_Q)\right), \tag{5}$$

where $\mathbf{D} \equiv ((B_1, S_1), (B_2, S_2), \ldots, (B_I, S_I))$ represents the numbers of classified *buys* and *sells* for days $i = 1, 2, \ldots, I$. Then, $\hat{\boldsymbol{\theta}}_Q = (\alpha_Q, \delta_Q, \mu_Q, \varepsilon_{b,Q}, \varepsilon_{s,Q}, q_Q)$ gives the consistent estimate of $\boldsymbol{\theta}_Q$ by MLE using (5). With $\hat{\boldsymbol{\theta}}_Q$ and $\varepsilon_Q \equiv \varepsilon_{b,Q} + \varepsilon_{s,Q}$, the PIN estimate is the ratio of mean informed trade to mean total trade as follows:

$$\text{PIN}_Q = \frac{\alpha_Q\mu_Q}{\alpha_Q\mu_Q + \varepsilon_Q}, \tag{6}$$

where the subscript Q indicate the estimate from Q-Method. Specifically, Q-Method denotes the above estimation procedure for PIN_Q.

Q-Method indirectly estimates the misclassification rate $(1-q)$ of the classification algorithm along with parameters of PIN Model or its extensions. Namely, the estimated misclassification rate is implied by a microstructure model, but is not directly obtained via examining *buys* and *sells*. Moreover, Q-Method guesses the arrival rates of B_i and S_i using the weighted sum of the arrival rates of actual $\widetilde{B_i}$ and $\widetilde{S_i}$ by q under each news type. Then, Q-Method uses the guessed arrival rates B_i and S_i to construe the likelihood of B_i and S_i, and generates the consistent estimates via MLE.

Moreover, the E-Method of [1] and [2] is a special case of the Q-Method with $q = 1$. Namely, they implicitly set $B_i = \widetilde{B_i}$

and $S_i = \widetilde{S_i}$, thus B_i and S_i are not affected by the misclassification. Therefore, with $q = 1$, the joint log-likelihood of observing a series of (B_i, S_i) over the past I trading days is the sum of the daily log-likelihoods $L(\boldsymbol{\theta}|B_i, S_i) \equiv \log(f(B_i, S_i|\boldsymbol{\theta}_Q = (\boldsymbol{\theta}, q = 1))$,

$$L(\boldsymbol{\theta}|\mathbf{D}) \equiv \sum_{i=1}^{I} L(\boldsymbol{\theta}|B_i, S_i). \quad (7)$$

With $\hat{\boldsymbol{\theta}} = (\alpha_E, \delta_E, \mu_E, \varepsilon_{b,E}, \varepsilon_{s,E})$, the estimate of $\boldsymbol{\theta}$ from MLE using (6), the PIN estimate is

$$\text{PIN}_E = \frac{\alpha_E \mu_E}{\alpha_E \mu_E + \varepsilon_E}, \quad (8)$$

where $\varepsilon_E \equiv \varepsilon_{b,E} + \varepsilon_{s,E}$. This implied assumption that $B_i = \widetilde{B_i}$ and $S_i = \widetilde{S_i}$ is unrealistic because of the ubiquitous problem of trade misclassification, and may lead to underestimation of the PIN. Moreover, the subscript E indicates the estimate from E-Method.

To sum up the above, Boehmer *et al*. in [10] argue that the misclassification reduces the discrepancy between the numbers of classified *buys* and *sells*. Thus, they conclude that the E-Method understates the true ratio of informed to uninformed trade arrival or the probability of an information event. Namely, PIN is understated. The Q-Method adjusts the arrival rates in a general case for mitigating the biases.

III. The Trade Classification Algorithms

The trade classification algorithms have been extensively used in microstructure studies. These algorithms usually infer the trade classification from trade and quote data by comparing the trade price to previous trade prices or to prevailing quotes. The two basic and common algorithms are the quote and tick methods.

The quote method uses the following criteria to sort trades: Trades with prices above the midpoint of bid and ask quotes, including those at the ask quote, are classified as buys; trades with price below the midpoint, including those at the bid quote, are classified as sells; and trades at the midpoint are left unclassified.

By contrast, the tick method sorts trades by comparing the price of the current trade to the price of the preceding trade. In the tick method, a trade with a price increase (decrease) relative to the previous trade price is an uptick (a downtick). A trade with the zero price change in which the last price change was an uptick (a downtick) is a zero-uptick (a zero-downtick). The tick method classifies upticks and zero-upticks as buys, and classifies downticks and zero-downticks as sells.

Incorporating the quote and tick methods, there are three algorithms proposed by Lee and Ready (LR) in [13], Eillis *et al.* (EMO) in [9] and Chakrabarty *et al.* (CBNV) in [8], respectively. LR uses the quote method to classify all trades possible, and then uses the tick method to sort the midpoint trades, which remain unclassified by the quote method. By contrast, EMO and CBNV first adopt the tick method to classify all trades, and then use the quote method further to refine some trades closed to bid or ask quotes. Specifically, EMO categorizes all trades executed at the ask quote as buys and all at the bid quotes as sells. All other trades are categorizes by the tick method. However, via dividing inside trades (for which the transaction is priced between bid and ask quotes) into deciles, Chakrabarty *et al.* in [8] in show that the quote method is better for trades with prices closer to the ask and the bid quotes and the tick rule does better when trade prices are closer to the midpoint. Therefore, CBNV method divides the spread into deciles (10% increments). Then the quote method is applied to the trades with prices within the three deciles from bid or ask quotes. Similarly, the others are classified by the tick method.

The prior studies [8] and [9] demonstrate that LR, EMO and CBNV methods result in statistically significant differences in the estimates of actual effective spreads and price impacts. This study examines whether the three algorithms result in the significant differences in the PIN estimates with Q- and E-Methods.

IV. Data and Estimation of PIN

Following criteria of [2], [6] and [18], this study selects the sample of common stocks listed on the New York Stock Exchange (NYSE) and American Stock Exchange (AMEX) for years 1993-2009. The selected sample excludes REITs, stocks of companies incorporated outside of the U.S., closed-end funds, and stocks with year-end price below $1. Also excluded are stocks in any year during which that did not have at least 60 days with quotes or trades, as PIN cannot be reliably estimated for such stock.

For estimating PIN, this study first retrieves transaction data from the Trade and Quote (TAQ) databases. This study then classifies trades as *buys* or *sells* via the LR, EMO and CBNV, respectively. Using *buys* and *sells* identified by each of the three algorithms, this study obtains PN estimates from Q- and E-Methods with the reformulated likelihood function derived from [19][1] and provided in Appendix. Hereafter, PIN_{QL}, PIN_{QE}, PIN_{QC}, PIN_{EL}, PIN_{EE}, and PIN_{EC} denote PIN estimates from Q- and E-Methods with LR, EMO and CBNV, respectively. The first subscript, *Q* or *E*, indicates the estimate from Q- or E-Methods; and the second subscript, *L*, *E* or *C*, indicates its data of *buys* and *sells* from LR, EMO or CBNV.

The Table I shows that the E- and Q-Methods generate the significantly different PIN estimates for each of the classification algorithms. The difference between E- and Q-Methods becomes greater in recent years. Namely, the misclassification may result in the biased PIN estimates especially for recent years. Moreover, E- or Q-Methods appear to generate the similar PIN estimates for variety algorithms because there are strong relationships among PIN_{EL}, PIN_{EE}, and PIN_{EC} as well as among PIN_{QL}, PIN_{QE}, and PIN_{QC}. Specifically, the untabulated result shows that he correlation confections among PIN_{EL}, PIN_{EE}, and PIN_{EC} are above 0.977 and these among PIN_{QL}, PIN_{QE}, and PIN_{QC} are

[1]Large daily numbers of *buys* and *sells* lead to the floating-point exception (FPE) in computing the likelihood function of PIN, and thus cause an underestimated PIN. Lin and Ke (2011) find such phenomenon and name it as FPE bias.

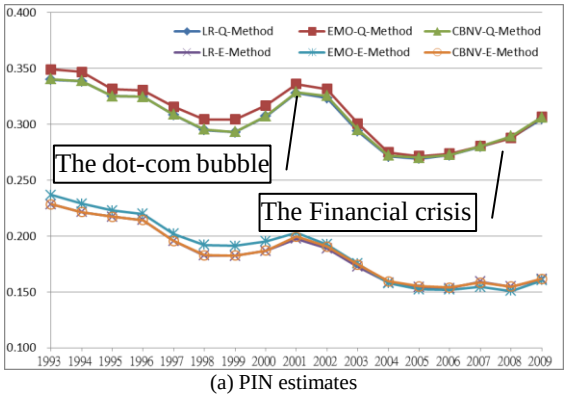
also above 0.976. Namely, the E- or Q-Methods are not sensitive to the algorithms.

TABLE I: THE SUMMARY OF PIN ESTIMATES

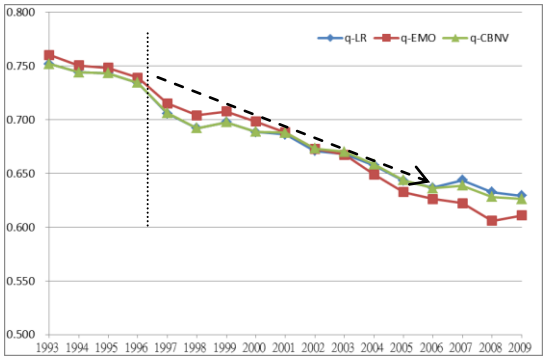
Year	N	LR			CBNV			EMO		
		PIN_{EL}	PIN_{QL}	T-test	PIN_{EE}	PIN_{QE}	T-test	PIN_{EC}	PIN_{QC}	T-test
1993	1584	0.23	0.34	-70.05***	0.24	0.35	-69.30***	0.23	0.34	-70.09***
1994	1626	0.23	0.34	-58.06***	0.24	0.35	-57.61***	0.23	0.34	-57.68***
1995	1687	0.22	0.34	-59.20***	0.23	0.35	-59.13***	0.22	0.34	-59.22***
1996	1713	0.22	0.33	-59.19***	0.22	0.33	-57.83***	0.22	0.33	-59.26***
1997	1742	0.21	0.33	-67.99***	0.22	0.33	-67.10***	0.21	0.33	-67.80***
1998	1802	0.20	0.31	-78.16***	0.20	0.32	-77.18***	0.20	0.31	-78.43***
1999	1818	0.18	0.30	-77.99***	0.19	0.30	-74.98***	0.18	0.30	-77.93***
2000	1778	0.18	0.29	-78.61***	0.19	0.30	-74.08***	0.18	0.29	-77.90***
2001	1640	0.19	0.31	-71.76***	0.20	0.32	-68.53***	0.19	0.31	-71.50***
2002	1580	0.20	0.33	-87.13***	0.20	0.34	-81.63***	0.20	0.33	-85.92***
2003	1571	0.19	0.32	-89.15***	0.19	0.33	-86.56***	0.19	0.33	-89.97***
2004	1583	0.17	0.29	-81.58***	0.18	0.30	-77.59***	0.17	0.30	-80.78***
2005	1572	0.16	0.27	-92.44***	0.16	0.28	-86.53***	0.16	0.27	-91.47***
2006	1565	0.15	0.27	-87.10***	0.15	0.27	-85.08***	0.16	0.27	-86.63***
2007	1558	0.15	0.27	-93.87***	0.15	0.27	-91.56***	0.15	0.27	-94.24***
2008	1480	0.16	0.28	-98.07***	0.15	0.28	-92.72***	0.16	0.28	-96.56***
2009	1358	0.16	0.29	-92.62***	0.15	0.29	-92.09***	0.16	0.29	-92.58***

TABLE II: THE SUMMARY OF ESTIMATES FOR CORRECT RATE Q

Year	N	q_{QL}	q_{QE}	q_{QC}	Year	N	q_{QL}	q_{QE}	q_{QC}
1993	1584	0.742	1993	1584	**2002**	1580	0.686	2002	1580
1994	1626	0.752	1994	1626	**2003**	1571	0.671	2003	1571
1995	1687	0.744	1995	1687	**2004**	1583	0.668	2004	1583
1996	1713	0.743	1996	1713	**2005**	1572	0.658	2005	1572
1997	1742	0.734	1997	1742	**2006**	1565	0.643	2006	1565
1998	1802	0.706	1998	1802	**2007**	1558	0.637	2007	1558
1999	1818	0.692	1999	1818	**2008**	1480	0.644	2008	1480
2000	1778	0.698	2000	1778	**2009**	1358	0.633	2009	1358
2001	1640	0.688	2001	1640					



(a) PIN estimates



(b) q estimates

Fig. 1. Yearly mean estimates of PIN and q.

With Fig. 1(a), we can get similar conclusions. Regardless of the algorithms, the mean PIN estimates from E-Methods have the close patterns, so do those from Q-Methods. Moreover, in contrast to the mean PIN estimates of E-Method, the mean PIN estimates of Q-Method are great and have significant improvements in the year of 2001, during which there is the dot-com bubble burst. Furthermore, after 2005 and during the Financial crisis of 2007–08, the mean PIN estimates of Q-Method have the increasing patterns but these of E-Method do not. Namely, Q-Method seems to mitigate the bias and generates a meaningful pattern.

This difference between E- and Q-Methods may be caused by the declining correct classification rates q. The Table II reports the means of q estimates for each year. Using these means, the Fig. 1(b) is plotted. With Fig. 1(b), the q estimates is significantly getting low after 1997. That is, the misclassification rate $(1 - q)$ is getting large and results in the great difference of PIN estimates between E- and Q-Methods. Moreover, for various algorithms, the Q-Method generates similar q estimates. However, the EMO algorithm appears to perform better than LR and CBNV before 2003, but it performs worse after 2003.

V. CONCLUSION

With this empirical result, the classification algorithms appear to have similar performances, and thus appear the close misclassification rates. Therefore, PIN estimates by E-Method are with trivial deviation and so do those by Q-Method. However, for the same algorithm, the large misclassification rate causes the substantial difference between the PIN estimates of E- and Q-Methods. Specially, in recent years, the misclassification is more serious, so the difference of estimates between E- and Q-Methods becomes greater.

This study is a beginning for further applications of Q-Method. With the empirical results, Q-Method helps reduce the PIN estimates bias caused by the trade misclassification. Therefore, the estimates of correct classification rate (q) may help to eliminate the overstatement of the effective spread caused by the misclassification in [9] and [20]. Specifically, the prior studies [9] and [20] adopt an effective spread calculated by 2×D×(Price – M) where D is the trade direction, +1 for a buy as well as –1 for a sell, and M mean the midpoint of bid-ask quotes. With the accurate rate q generated by the Q-Method, the revised effective spread may be

$$\begin{aligned} &2(q\mathrm{D} + (1-q)(-\mathrm{D}))\times(\mathrm{Price} - \mathrm{M}) \\ &= 2(2q-1)\times\mathrm{D}\times(\mathrm{Price} - \mathrm{M}). \end{aligned} \tag{7}$$

APPENDIX

Let $\mu_c = q\mu$, $\mu_{\underline{c}} = (1-q)\mu$, $\lambda_{Nb} = q\varepsilon_b + (1-q)\varepsilon_s$, $\lambda_{Ns} = q\varepsilon_s + (1-q)\varepsilon_b$, $\lambda_{Bb} = \lambda_{Nb} + \mu_{\underline{c}}$, $\lambda_{Bs} = \lambda_{Ns} + \mu_c$, $\lambda_{Gb} = \lambda_{Nb} + \mu_c$, and $\lambda_{Gs} = \lambda_{Ns} + \mu_{\underline{c}}$.

Then, the daily log-likelihood function with computing

stability is as follows:

$$L_A(\boldsymbol{\theta}_Q|B_i, S_i) \equiv \log(f(B_i, S_i|\boldsymbol{\theta}_Q)) = \log(\exp(e_{1,i} - e_{max,i}) + \exp(e_{2,i} - e_{max,i}) + \text{epx}(e_{3,i} - e_{max,i})) + e_{max,i} + B_i \log(\lambda_{Nb}) + S_i \log(\lambda_{Ns}) - (\varepsilon_b + \varepsilon_s + \mu) - \log(S_i!B_i!), \quad (8)$$

where $e_{1,i} = \log(\alpha\delta) + B_i\log(1+\mu_c / \lambda_{Nb}) + S_i\log(1+\mu_c / \lambda_{Ns})$; $e_{2,i} = \log(\alpha(1-\delta)) + B_i\log(1+\mu_c / \lambda_{Nb}) + S_i\log(1+ \mu_c / \lambda_{Ns})$; $e_{3,i} = \log(1-\alpha) + \mu$; $e_{max,i} = \max(e_{1,i}, e_{2,i}, e_{3,i})$ and the term $\log(S_i!B_i!)$ is dropped in computing.

Therefore, the joint log-likelihood of observing $\mathbf{D} \equiv ((B_1, S_1), (B_2, S_2), \ldots, (B_I, S_I))$, $L_A(\boldsymbol{\theta}_Q|\mathbf{D}) \equiv \sum_{i=1}^{I} L_A(\boldsymbol{\theta}_Q| B_i, S_i)$ is adopted in the PIN estimation.

REFERENCES

[1] D. Easley, N. M. Kiefer, M. O'Hara, and J. Paperman, "Liquidity, information, and infrequently traded stocks," *Journal of Finance*, vol. 51, pp. 1405–1436, 1996.

[2] D. Easley, S. Hvidkjaer, and M. O'Hara, "Is information risk a determinant of asset returns?" *Journal of Finance*, vol. 57, pp. 2185–2221, 2002.

[3] A. Ascioglu, S. P. Hegde, and J. B. McDermott, "Information asymmetry and investment–cash flow sensitivity," *Journal of Banking & Finance*, vol. 32, pp. 1036–1048, 2008.

[4] P. Brockman and X. Yan, "Block ownership and firm-specific information," *Journal of Banking & Finance*, vol. 33, pp. 308–316, 2009

[5] A. Boulatov, B. C. Hatch *et al*., "Dealer attention, the speed of quote adjustment to information, and net dealer revenue," *Journal of Banking & Finance*, vol. 33, pp. 1531–1542, 2009.

[6] D. Easley, S. Hvidkjaer, and M. O'Hara, "Factoring information into returns," *Journal of Financial and Quantitative Analysis*, vol. 45, no. 2, pp. 293–309, 2010.

[7] D. Easley, M. O'Hara, and J. Paperman, "Financial analysts and information-based trade," *Journal of Financial Markets*, vol. 1, no. 2, pp. 175–201, 1998.

[8] B. Chakrabarty, B. Li, V. Nguyen, and R. A. Van Ness, "Trade classification algorithms for electronic communications network trades," *Journal of Banking & Finance*, vol. 31, pp. 3806–3821, 2007.

[9] K. Ellis, R. Michaely, and M. O'Hara, "The accuracy of trade classification rules: Evidence from Nasdaq," *Journal of Financial and Quantitative Analysis*, vol. 35, no. 4, pp. 529–551, 2000.

[10] E. Boehmer, J. Grammig, and E. Theissen, "Estimating the probability of informed trading – does trade misclassification matter?" *Journal of Financial Markets*, vol. 10, pp. 26–47, 2007.

[11] M. J. Barclay, T. Hendershott, and D. T. McCormick, "Competition among trading venues: Information and trading on electronic communication networks," *Journal of Finance*, vol. 58, no. 6, pp. 2365–2637, 2003.

[12] T. Hendershott and C. M. Jones, "Island goes dark: transparency, fragmentation, and regulation," *Review of Financial Studies*, vol. 18, pp. 743–793, 2005.

[13] C. Lee and M. Ready, "Inferring trade direction from intraday data," *Journal of Finance*, vol. 46, pp. 733–746, 1991.

[14] C. Lee and B. Radhakrishna, "Inferring investor behavior: evidence from TORQ," *Journal of Financial Markets*, vol. 3, no. 2, pp. 83–111, 2000.

[15] E. R. Odders-White, "On the occurrence and consequences of inaccurate trade classification," *Journal of Financial Markets*, vol. 3, no. 3, pp. 259–286, 2000.

[16] J. Grammig, D. Schiereck, and E. Theissen, "Knowing me, knowing you: Trader anonymity and informed trading in parallel markets," *Journal of Financial Markets*, vol. 4, no. 4, pp. 385–412, 2001.

[17] J. Duarte and L. Young, "Why is PIN priced?" *Journal of Financial Economics*, vol. 91, pp. 119–138, 2009.

[18] P. Mohanram and S. Rajgopal, "Is PIN priced risk?" *Journal of Accounting and Economics*, vol. 47, pp. 226-243, 2009.

[19] H.-W. Lin and W.-C. Ke, "A computing bias in estimating the probability of informed trading," *Journal of Financial Markets*, vol. 14, no. 4, pp. 625-640, 2011.

[20] M. Peterson and E. Sirri, "Evaluation of the biases in execution cost estimation using trade and quote data," *Journal of Financial Markets*, vol. 6, pp. 259–280, 2003.

Wen-Chyan Ke is an assistant professor in the Department of Finance and Cooperative Management, National Taipei University, Taiwan. He received his Ph.D. degree in 2009 from National Taiwan University. His research interests are option pricing, market microstructure and neural networks. His recent work has been published in *Journal of Financial Markets* and *Computational Economics*.