Test of PIN Algorithms through Simulation

Marek Jochec*
School of Economics, Technical University Ostrava
Nazarbayev University

Abstract

I use simulated series of buys and sells to test nine different PIN estimation algorithms implemented in SAS, proc NLP. I conclude that the algorithms do a good job estimating PIN, on average, when applied to series generated from high-PIN data generating process (PIN>0.2). However, when applied to series generated from zero-PIN or low-PIN (PIN<0.2) data generation process, the algorithms yield overstated PIN estimates. Supplying initial values to parameters in the estimation has strong effect on the estimated PIN. Factorization of the likelihood function plays a less important role. Not supplying initial values dramatically overstates PINs for zero-PIN and low-PIN data: zero-PIN data shows PIN as high as 0.26; low-PIN data tend to have PINs overstated by 0.05 to 0.10. Supplying initial values as in Lin and Ke (2011) eliminates the bias for low-PIN data but not for zero-PIN data. Supplying initial values which correspond to the null hypothesis (of no informed trading) yields zero PIN estimates for low-Pin as well as high-PIN data.

JEL classification: G12; G14; C13; C63
Key words: PIN, probability of informed trading, market microstructure, simulation

* E-mail address: marek.jochec@nu.edu.kz.

Introduction

PIN (=probability of informed trading) is a widely-used market microstructure measure. The concept was developed by Easley, Kiefer, O'Hara and Paperman (1996) and Easley, Hvidkjaer, and O'Hara (2002). They propose a structural model in which uninformed liquidity traders submit their market orders (to buy or sell a stock) randomly, at a steady rate, throughout some period of time. The daily arrival rates of buy-initiated trades (sell-initiated trades) are ε_b (ε_s) and are the means of two independent Poisson processes. On some days (with probability α), private information becomes revealed to informed traders and they trade on it, submitting their trades at a rate μ, also a mean of an independent Poisson process. If the information is "good" (with probability δ), the informed trades will be on the buy side; if "bad" (with probability 1-δ), the informed trades will be on the sell side. Thus, on days with no private information event, the expected numbers of daily buys and sells will be ε_b and ε_s. On days with private information event, the average numbers will be $\mu+\varepsilon_b$ and ε_s if the information is "good"; and ε_b and $\mu+\varepsilon_s$, if the information is "bad" (see Figure 1.). PIN is then defined as the ratio of the number of informed trades to the number of all trades, informed and uninformed, over a period of time: $PIN = \alpha\mu / (\alpha\mu+\varepsilon_b+\varepsilon_s)$. By construction, PIN is a unit-less number between zero and one, and thus can be interpreted as a probability. PIN is usually estimated for stock years or stock quarters. It is used in many studies as an independent variable measuring the extent to which a particular stock, in particular period of time, is likely to be subjected to informed trading. Possibility of the presence of informed trader constitutes a risk to uninformed liquidity traders, since, by definition, they will be on the wrong side of the transaction. This additional risk may effect stock returns, bid-ask spreads etc. (For example, see the studies mentioned in Benos and Jochec (2007) or Brennan, Huh and Subrahmanyam (2013)).

The PIN model can be used both for estimating the parameters α, δ, μ, ε_b, and ε_s from the buy and sell data, such as for simulating the random buy and sell data on a computer.

In the estimation, a series of observed daily numbers of buys and sells serves as an input for the estimation of the unobserved most likely parameters α, δ, μ, ε_b, and ε_s by using maximum likelihood estimation technique (MLE). The MLE estimates can be obtained, for example, using SAS proc NLP, STATA ml procedure, Matlab, or Microsoft Excel Solver.

Specifically, the likelihood function for a single day, on which *B* buys and *S* sells were observed, is given by:

$$
\begin{aligned}
L(\theta|B,S) = &(1-\alpha)e^{-\epsilon_b}\frac{\epsilon_b^B}{B!}e^{-\epsilon_s}\frac{\epsilon_s^S}{S!} \\
&+ \alpha\delta e^{-\epsilon_b}\frac{\epsilon_b^B}{B!}e^{-(\mu+\epsilon_s)}\frac{(\mu+\epsilon_s)^S}{S!} \\
&+ \alpha(1-\delta)e^{-(\mu+\epsilon_b)}\frac{(\mu+\epsilon_b)^B}{B!}e^{-\epsilon_s}\frac{\epsilon_s^S}{S!}
\end{aligned}
$$

where θ = (α, δ, μ, ε_b, ε_s) is the parameter vector.

Since the information events, the character of the news (good/bad), and arrival rates of orders (uninformed and informed) are independent from each other and across days, the likelihood function for a period of *I* days has the form:

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

where (B_i, S_i) are the day *i* trades and M = ((B_i, S_i), ... , (B_I, S_I)) is the series of daily numbers of buys and sells. Maximizing the last expression over θ is the way to find the individual parameters of the PIN expression and thus to calculate PIN. (For more details refer to, for example, Yan and Zhang (2012), section 2.)

In the simulation, the random numbers of daily buys and sells for a stock with certain PIN are obtained from a DGP (data generating process) with the parameters α, δ, μ, ε_b, and ε_s. The DGP is in the form of a probability tree with the respective Poisson processes at each final node. The number of buys and sells observed on a particular day is the result of random draws of 0 or 1, drawn from distributions with means α and δ (this determines at which final node of the probability tree we end up), and of random draws from Poisson distributions with means ε_b, ε_s, and possibly μ.

Early studies about PIN (or using PIN) did not worry too much about the details of numerical estimation techniques and algorithms used to obtain PIN MLE estimates. Optimization procedure was assumed to search through different parameter vectors α, δ, μ, ε_b, ε_s over the whole parameter space (which is defined as follows: α and δ are real numbers from zero to one, μ, ε_b, ε_s are integers from zero to infinity) and output the one that maximizes the likelihood function.

Lin and Ke (2011) showed that, when a numerical maximization of the likelihood function is performed on a typical computer, the floating-point exception (which happens when computer encounters a number bigger than (about) 10^{308} or smaller than 10^{-307}) "may eliminate feasible solutions to the actual parameters in the optimization problem". In other words, the computer will try to find the maximizing parameter vector from a subset of parameter space which may not include the globally-maximizing parameter vector. The parameter vector found in this way gives biased PIN. They propose an improved factorization of the likelihood function ("LK factorization"), which they show is performing better than previously used factorizations ("EHO 2002" and "EHO 2010"). (By "factorization" we mean the way the likelihood function is written algebraically. In the absence of hardware computing limitations, each factorization would be evaluated equally for any set of parameters supplied and would give the same maximum likelihood parameter estimates. But because of hardware limitations resulting in overflow or underflow, the different factorizations may give different results.)

In related work, Yan and Zhang (2012) show the effect of the choice of (or absence of) the initial values to parameters supplied to the optimization module. (Initial values can be supplied in SAS proc NLP with INIT option. If they are not supplied, the procedure starts with initial values obtained by its own method.) They further improve the estimation method by running the optimization up to 125 (or 729) times, each time with different initial parameter values. They choose the values in such a way that they better explore the parameter space, and thus find solutions which could be otherwise missed. They also show that some solutions may be the boundary solutions with respect to α or δ (i.e., solutions in which α or δ is 0 or 1). They argue that such solutions are implausible and further adjust the method, using the following rule: If all 125 sets of initial values give a boundary solution, they accept the one giving the highest likelihood. If some non-boundary solutions are found, they choose the one giving the highest likelihood among them.

Studies like Lin and Ke (2011) or Yan and Zhang (2012) show that it is possible that for the same buy and sell data, the computer may give different PIN estimates, based on the factorization used and the choice of initial values. If one does not pay attention to this, the chances are some estimated PINs will be biased.

In this study, I test how the choice of the factorization and initial parameters affects the estimated PINs. I do this by simulating random buy and sell data on a computer and then estimating the PINs. The data comes from a known DGP (I choose DGP parameters α, δ, μ, ε_b, ε_s) and thus know the implied PIN of the generated data. I can then compare the estimated PINs with the implied PIN. The generated data is random and thus the estimated individual PINs will usually differ from the implied PIN. However, if the estimation method works well, the *average* estimated PIN, over a sufficiently large sample, should equal the implied PIN. The large difference between average estimated PIN and implied PIN suggests a systematic weakness of the estimation technique.

Methodology

I compare three possible factorizations of the likelihood function: EHO 2002, EHO 2010, and LK 2011, each with three ways of choosing the initial values: the initial values are supplied and are equal to the parameters used in the DGP, the initial values are not supplied, and 125 sets of initial values are supplied as in Yan and Zhang (2012). This gives me the comparison of nine PIN estimation algorithms.

I start by generating 1,000 random series of signed trades. Each series has 60 data points, corresponding to 60 trading days (a quarter of a year). Each data point is a pair of two numbers: the number of buys, B, and the number of sells, S, on that day. I use the structural model which underlies the PIN as the data generation process. Thus, different parameter vectors (α, δ, μ, ε_b, ε_s) give different random series. I then estimate PIN of the generated series, and obtain 1,000 PIN estimates for each series. Appendix A provides the Excel VBA code used to generate the data.

I choose uninformed B (daily number of buys) and S (daily number of sells) to follow independent Poisson distribution with Poisson's $\lambda = 30$ ($\varepsilon_s = \varepsilon_b = \varepsilon = 30$). By assuming the buy-sell symmetry, I decrease the number of parameters to four. Then I generate informed trades from Poisson distribution with Poisson $\lambda = \mu = 15, 30, 45,$ and 60, and add them to uninformed B and S numbers. μ is chosen such that, with combination of chosen α, I get implied PINs from 0 to 0.47. I take $\varepsilon = 30$ because this is a round number close to the actual average of buys and sells for a stocks in our 2007 sample. The absolute value of ε does not matter: it is only the ratio of μ to ε that matters. In my case, I take μ equal to $\frac{1}{2}\varepsilon$, ε, $1\frac{1}{2}\varepsilon$ and 2ε. The informed trades are added to the uninformed trades if a random draw from continuous uniform distribution on zero-one interval is less than α. They are added to the buys (sells) if a random draw from continuous uniform distribution on zero-one interval is less than (more than) δ. I choose α to be either 0, 0.3, 0.6, or 0.9 (to span the zero-one probability space), and δ equal to either 0 or 0.5 ($\delta = 1$ is equivalent to $\delta = 0$ due to symmetry of buys and sells). (I also tried delta = .25, with no effect on the results.) In total, I obtain 25 parameter vectors: The first gives me one thousand series of 60 uninformed B,S pairs; the remaining 24 the same series but increased for informed sales according to the above algorithm.

For example, consider a DGP with $\alpha = 0.6$ (there is private news on 60% of the days), delta = 0 (the news is always negative), $\mu = 45$ (on average, 45 informed trades are placed on days with private news). Since delta is 0, all the informed trades are added to S. The data generated in this way should have average PIN equal to .6 x 45 ÷ (.6 x 45 + 30 + 30) = .31.

After I have generated the data for all 25 specifications, I estimate PIN of each individual B,S series, and obtain their averages, standard deviations, and histograms.

I estimate PINs using three different factorizations: (1) EHO 2002, (2) EHO 2010, and (3) LK (Lin Ke 2011). All are run with Newton Raphson technique (in SAS, option "`tech=newrap`"). Appendix B provides the SAS code. For each factorization, I consider three versions of initial parameters: DGPinit uses the parameter vector used for data generation as initial parameters for PIN estimation. Noinit is with no initial parameters supplied. YZinit is using Yan and Zhang (2012) method to use up to 125 initial parameter vectors and selecting the non-boundary solution with the highest likelihood, or, if no non-boundary solution exists, then boundary solution with the highest likelihood. In total, I obtain nine sets of simulation results.

Results

The average estimated PIN for the nine tested algorithms and 25 parameter vector specifications is reported in Table 1. The highlighted cells in the last nine columns show average PINs which differ from theoretical PIN by more than 0.02. If the DGP parameters are used as the initial values, the average PIN is estimated precisely by all three factorizations. In the next six columns there are quite a few PINs differing from the theoretical PIN by more than 0.02. The misestimated PINs are

concentrated near the top of the table among low-PIN specifications. The only method which estimates PINs well is LK factorization with YZ initial parameters. Even this method overestimates zero-PIN data by 0.06.

Not providing initial parameters clearly leads to overestimation of zero or low PINs (columns under NO-init heading). Previous studies which relied on PIN and used PIN estimated without the provision of initial values are most likely biased, no matter which factorization method was used.

Next I explore if the estimation can be improved by relaxing the non-boundary solution requirement in YZ method. I speculate that zero-PIN data has PINs overestimated because I use algorithm that preferably picks the non-boundary solution, if one exists (while data is generated with $\alpha = 0$). I rerun the estimation, relaxing the non-boundary solution preference. It turned out that this was indeed the case: not preferring non-boundary solution brings average estimated pin for all three factorizations back to theoretical pin (except for the zero-PIN data in model 1, where average estimated PIN is now 0.06 for all three factorizations). I conclude that relaxing the preference for non-boundary solution improves the EHO 2002 and EHO 2010 method and bring them at par with LK method. One should note that the positive average PIN estimated from zero-PIN generated data can be simply a result of PIN being a number bounded at zero. Since there cannot be negative PINs, an average taken over some zero PINs and some positive PINs will be a positive number. (In other words, to get zero average PIN, all PINs would have to be estimated at zero.) It looks like some random series of data generated with no informed trading received such B and S values that they look as if at least some informed trading is present. I.e., they resemble positive-PIN data rather than zero-PIN data.

Next I look how the individual parameters α, δ, μ, ε_b, and ε_s are estimated. If the parameters used in the DGP are used as the initial parameters in the estimation, all parameters are estimated correctly (not reported). Tables 3A and 3B show how the individual parameters are estimated if either no or YZ initial values are used. One notable finding is that zero-PIN data give positive PIN estimates by overestimating μ and α, in all six specifications.

For comparison, I rerun the tests with three more ways to choose initial values to estimated parameters. The first way chooses "null hypothesis" initial values (parameters consistent with no informed trading): I use average number of buys (sells) as initial value for ε_b (ε_s), and initial value for α, δ, and μ is zero. Second way uses random initial values: α, δ are random numbers on zero-one interval, ε_b, ε_s, and μ are random integers ranging from 0 to the (global) average of (B+S). The third way uses high-PIN initial values: $\alpha = 1$, μ = average (B+S)/2, ε_b = average B/2, and ε_s = average S/2. The results are presented in Table 4. Interesting result is that the null-hypothesis initial values give zero PIN estimates for all data. Basically, the initial values are accepted as maximum-likelihood parameters. Table 4 shows the results.

Finally, Table 5 shows the standard deviations of PINs obtained from all PIN estimation algorithms considered in this study. This gives the idea how much dispersed the individual PIN estimates tend to be if estimated by different algorithms.

Conclusion

I test several PIN estimation algorithms through simulation.

First, I generate random series of buys and sells numbers, where I use the probability tree assumed by EHO 2002 structural model of informed trading as the tree of the data generating process. By choosing the values of the parameters of the simulation, I obtain number of series of buys and sells data with known implied PIN. I generate data without informed trading (zero PIN), with some informed trading (low PIN), and with intense informed trading (high PIN).

Then I use the generated data to test different PIN estimation algorithms. Algorithms differ by the factorization of the likelihood function, and by the choice (or no choice) of the initial values of the

estimated parameters. I conclude that the algorithms do a good job estimating PIN, on average, when applied to series generated from high-PIN data generating process (PIN>0.2). However, when applied to series generated from zero-PIN or low-PIN (PIN<0.2) data generation process, the algorithms yield overstated PIN estimates. Not supplying initial values dramatically overstates PINs for zero-PIN and low-PIN data: zero-PIN data shows PIN as high as 0.26; low-PIN data tend to have PINs overstated by 0.05 to 0.10. Supplying initial values as in Lin and Ke (2011) eliminates the bias for low-PIN data but not for zero-PIN data. Supplying initial values which correspond to the null hypothesis (of no informed trading) yields PIN estimates close to zero, even for high-PIN data.

The PIN model is imposed on data by the researcher; the null hypothesis should assume the PIN is zero and there are no informed traders. Many studies before LK and YZ algorithms, such as Benos Jochec (2007), report average estimated PIN to be in the range of 0.20 to 0.25. This is the same as average PIN estimated from generated zero-PIN data. In other words, it is possible that positive PIN is the result of the estimation method rather than of informed trading.

A possible remedy to this problem is to rerun the previous studies, but conservatively replacing all low PINs with zero. A possible rule of thumb can be to take PIN to be zero whenever it is estimated at 0.25 or less (.06, .10, depending on the estimation method used in that study). This rule is based on the average estimated PIN for zero-PIN stocks in this study. (One could also base the rule on the Nth percentile, if an informed choice about N can be made.) My simulation also suggests that high-PIN stocks have their PINs estimated correctly irrespective of the estimation method.

References

Benninga Simon, "Financial Modeling", 3rd edition, The MIT Press, ISBN 978-0-262-02628-4

Benos Evangelos and Marek Jochec, 2007, "Testing the PIN Variable", working paper, Nazarbayev University

Brennan Michael, Sahn-Wook Huh, and Avanidhar Subrahmanyam, 2013, "Informed Trading and the Pricing of Good and Bad Information in the Cross-Section of Expected Stock Returns", working paper

Easley David, Soeren Hvidkjaer and Maureen O'Hara, 2002, "Is Information Risk a Determinant of Asset Returns?", Journal of Finance 47, 2185-2221

Easley David, Nicholas M. Kiefer, Maureen O'Hara and Joseph B. Paperman, 1996, "Liquidity, Information, and Infrequently Traded Stocks", Journal of Finance 51, 1405-1436

Lin Hsiou-Wei Lin and Wen-Chyan Ke, 2011, "A Computing Bias in Estimating the Probability of Informed Trading", Journal of Financial Markets 14, 625-640

Yan Yuxing and Shaojun Zhang, 2012, "An Improved Estimation Method and Empirical Properties of the Probability of Informed Trading", Journal of Banking and Finance 36, 454-467

Figure 1: Probability tree of the assumed structure of the trading process underlying PIN (EHO 2002).

Signal
High δ

Buy Arrival Rate:
$\varepsilon_b + \mu$

Sell Arrival Rate:
ε_s

Information Event
Occurs α

Signal Low
$(1 - \delta)$

Buy Arrival Rate:
ε_b

Sell Arrival Rate:
$\varepsilon_s + \mu$

Information Event Does Not
Occur $(1 - \alpha)$

Buy Arrival Rate:
ε_b

Sell Arrival Rate:
ε_s

Table 1:

Average PIN estimated from 1,000 series of 60 B,S pairs randomly generated from the data generation process assumed by the PIN model with parameters α, δ, μ, ε. The parameter α is a probability that there is an information event on a given day, δ is a probability that the news is positive, μ is expected number of informed trades on a day with information event, ε is expected number of uninformed buys on any day. Expected number of uninformed sells is assumed to be the same as the expected number of uninformed buys. Parameters are chosen such that PIN implied by the DGP ranges from 0 to 0.47. DGP-init supplies the DGP parameters as the initial parameters in the estimation, NO-init does not supply any initial parameters, YZ-init supplies the initial parameters calculated as in Yan Zhang 2012. For each case of initial parameters, three factorizations of the likelihood functions are used: EHO 2002, EHO 2010, and LK. The SAS code for each factorization is provided in Appendix B. (The models are sorted by implied PIN, μ, and δ.)

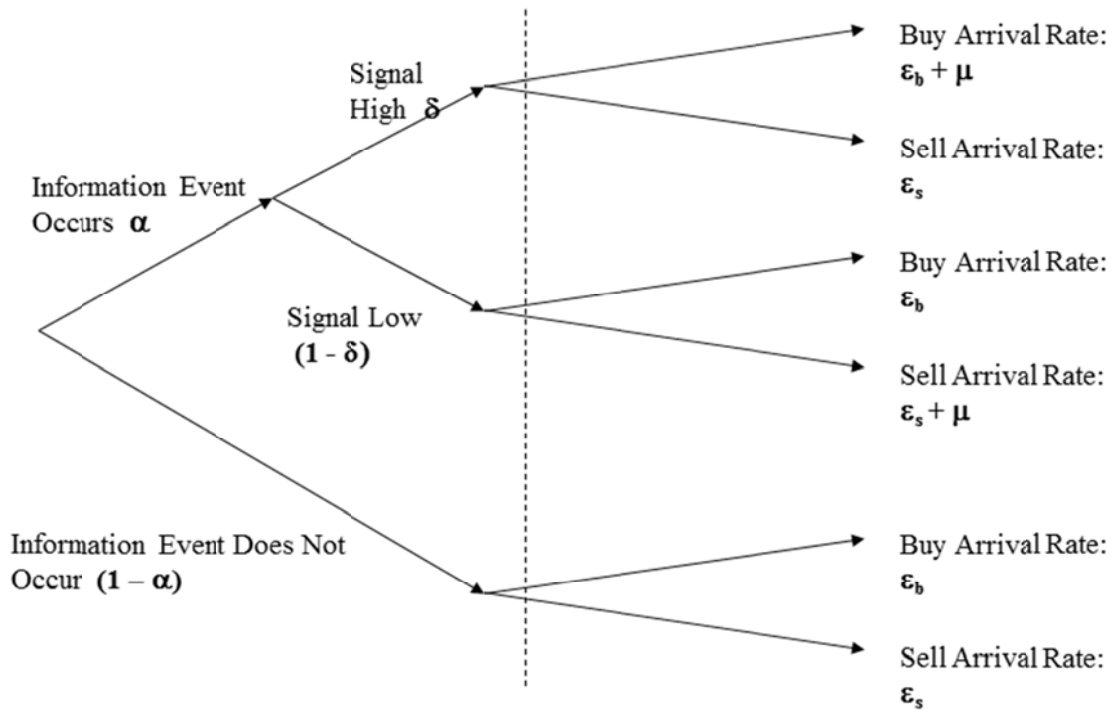| model	parameters of DGP					DGP-init			NO-init			YZ-init		
					implied	PIN	PIN	PIN	PIN	PIN	PIN	PIN	PIN	PIN
	ε	μ	α	δ	PIN	EHO2002	EHO2010	LK	EHO2002	EHO2010	LK	EHO2002	EHO2010	LK
1	30	0	0.0	0.0	**0.00**	0.00	0.00	0.00	***0.25***	***0.23***	***0.26***	***0.10***	***0.10***	***0.06***
2	30	15	0.3	0.0	**0.07**	0.07	0.07	0.07	***0.17***	***0.15***	***0.14***	***0.14***	***0.14***	0.08
3	30	15	0.3	0.5	**0.07**	0.07	0.07	0.07	***0.13***	***0.13***	***0.14***	0.07	0.07	0.07
4	30	15	0.6	0.0	**0.13**	0.13	0.13	0.13	***0.18***	***0.18***	***0.20***	***0.23***	***0.23***	0.13
5	30	15	0.6	0.5	**0.13**	0.13	0.13	0.13	0.15	0.14	0.14	0.13	0.13	0.13
6	30	30	0.3	0.0	**0.13**	0.13	0.13	0.13	0.14	0.15	0.15	***0.21***	***0.20***	0.14
7	30	30	0.3	0.5	**0.13**	0.13	0.13	0.13	0.14	0.14	0.14	0.13	0.13	0.13
8	30	15	0.9	0.0	**0.18**	0.18	0.18	0.18	***0.28***	***0.25***	***0.25***	***0.27***	***0.27***	0.17
9	30	15	0.9	0.5	**0.18**	0.18	0.18	0.18	0.19	0.20	0.20	0.19	0.19	0.18
10	30	45	0.3	0.0	**0.18**	0.18	0.18	0.18	0.19	0.20	0.19	***0.25***	***0.25***	0.19
11	30	45	0.3	0.5	**0.18**	0.18	0.18	0.18	0.19	0.19	0.19	0.18	0.18	0.18
12	30	30	0.6	0.0	**0.23**	0.23	0.23	0.23	0.24	0.23	0.25	0.24	0.24	0.23
13	30	30	0.6	0.5	**0.23**	0.23	0.23	0.23	0.24	0.23	0.25	0.23	0.23	0.23
14	30	60	0.3	0.0	**0.23**	0.23	0.23	0.23	0.24	0.23	0.23	0.24	0.24	0.23
15	30	60	0.3	0.5	**0.23**	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
16	30	30	0.9	0.0	**0.31**	0.31	0.31	0.31	0.32	0.31	0.32	***0.37***	***0.38***	0.31
17	30	30	0.9	0.5	**0.31**	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
18	30	45	0.6	0.0	**0.31**	0.31	0.31	0.31	0.31	0.31	0.31	0.33	0.33	0.31
19	30	45	0.6	0.5	**0.31**	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
20	30	60	0.6	0.0	**0.38**	0.37	0.37	0.37	0.38	0.37	0.38	0.38	0.38	0.37
21	30	60	0.6	0.5	**0.38**	0.37	0.37	0.37	0.38	0.38	0.37	0.37	0.37	0.37
22	30	45	0.9	0.0	**0.40**	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
23	30	45	0.9	0.5	**0.40**	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
24	30	60	0.9	0.0	**0.47**	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47
25	30	60	0.9	0.5	**0.47**	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47

Table 2.:

The difference between average estimated PIN and the PIN implied by the data generation process. Only PINs which differ by more than 0.02 are shown. All differences are positive, meaning that the algorithms overestimate the PIN. The biggest bias is for zero-PIN data, where all estimation techniques overestimate PIN. The bias ranges from 0.26 (LK factorization with no initial parameters) to 0.06 (LK factorization with YZ initial parameters). There are no cases of PIN being underestimated.

model	parameters of DGP					NO-init			YZ-init		
					implied	PIN	PIN	PIN	PIN	PIN	PIN
	ε	μ	α	δ	PIN	EHO2002	EHO2010	LK	EHO2002	EHO2010	LK
1	30	0	0.0	0.0	0.00	0.25	0.23	0.26	0.10	0.10	0.06
2	30	15	0.3	0.0	0.07	0.10	0.08	0.07	0.07	0.07	
3	30	15	0.3	0.5	0.07	0.06	0.06	0.07			
4	30	15	0.6	0.0	0.13	0.05	0.05	0.07	0.10	0.10	
5	30	15	0.6	0.5	0.13						
6	30	30	0.3	0.0	0.13				0.08	0.07	
7	30	30	0.3	0.5	0.13						
8	30	15	0.9	0.0	0.18	0.10	0.07	0.07	0.09	0.09	
9	30	15	0.9	0.5	0.18						
10	30	45	0.3	0.0	0.18				0.07	0.07	
11	30	45	0.3	0.5	0.18						
12	30	30	0.6	0.0	0.23						
13	30	30	0.6	0.5	0.23						
14	30	60	0.3	0.0	0.23						
15	30	60	0.3	0.5	0.23						
16	30	30	0.9	0.0	0.31				0.06	0.07	
17	30	30	0.9	0.5	0.31						
18	30	45	0.6	0.0	0.31						
19	30	45	0.6	0.5	0.31						
20	30	60	0.6	0.0	0.38						
21	30	60	0.6	0.5	0.38						
22	30	45	0.9	0.0	0.40						
23	30	45	0.9	0.5	0.40						
24	30	60	0.9	0.0	0.47						
25	30	60	0.9	0.5	0.47						

Table 3A.

Average α, δ, μ, ε_b, ε_s and PIN estimated from 1,000 series of 60 B,S pairs randomly generated from the data generation process assumed by the PIN model with parameters α, δ, μ, $\varepsilon=\varepsilon_b=\varepsilon_s$, if no initial values for parameters α, δ, μ, ε_b, ε_s are used. The parameter α is a probability that there is an information event on a given day, δ is a probability that the news is positive, μ is expected number of informed trades on a day with information event, ε_b is expected number of uninformed buys on any day, and ε_s is expected number of uninformed sells on any day. Expected number of uninformed sells is assumed to be the same as the expected number of uninformed buys in the data generation. Parameters are chosen such that PIN implied by the DGP ranges from 0 to 0.47. The factorization used is EHO 2002, EHO 2010, and LK. (The models are sorted by implied PIN, μ, and δ.) "isb" variable shows the percentage of boundary solutions (in which α or δ is equal to zero or one).

model	parameters of DGP					EHO 2002						EHO 2010						LK					
	ε	μ	α	δ	PIN	ε_b	ε_s	μ	α	δ	PIN	ε_b	ε_s	μ	α	δ	PIN	ε_b	ε_s	μ	α	δ	PIN
1	30	0	0.0	0.0	**0.00**	25	20	15	0.98	0.65	0.25	26	20	14	0.97	0.65	0.23	25	19	16	0.99	0.69	0.26
2	30	15	0.3	0.0	**0.07**	24	29	17	0.61	0.08	0.17	27	27	16	0.60	0.18	0.15	28	27	16	0.58	0.19	0.14
3	30	15	0.3	0.5	**0.07**	28	28	16	0.52	0.48	0.13	28	28	16	0.52	0.48	0.13	28	28	16	0.54	0.49	0.14
4	30	15	0.6	0.0	**0.13**	28	29	17	0.69	0.07	0.18	27	30	17	0.68	0.03	0.18	29	26	18	0.74	0.19	0.20
5	30	15	0.6	0.5	**0.13**	29	30	16	0.64	0.48	0.15	30	30	15	0.63	0.49	0.14	29	30	15	0.63	0.49	0.14
6	30	30	0.3	0.0	**0.13**	30	29	29	0.34	0.03	0.14	29	30	29	0.36	0.02	0.15	29	30	29	0.36	0.02	0.15
7	30	30	0.3	0.5	**0.13**	30	30	30	0.32	0.50	0.14	29	30	30	0.32	0.50	0.14	30	30	30	0.32	0.50	0.14
8	30	15	0.9	0.0	**0.18**	23	30	22	0.93	0.01	0.28	27	28	19	0.93	0.11	0.25	25	29	20	0.92	0.02	0.25
9	30	15	0.9	0.5	**0.18**	30	30	16	0.90	0.50	0.19	30	28	16	0.91	0.53	0.20	29	30	16	0.90	0.49	0.20
10	30	45	0.3	0.0	**0.18**	29	29	44	0.33	0.02	0.19	28	28	42	0.37	0.08	0.20	29	29	43	0.35	0.04	0.19
11	30	45	0.3	0.5	**0.18**	29	29	44	0.32	0.49	0.19	30	30	45	0.31	0.49	0.19	29	29	44	0.33	0.48	0.19
12	30	30	0.6	0.0	**0.23**	29	29	30	0.63	0.03	0.24	29	29	29	0.61	0.02	0.23	28	29	29	0.64	0.05	0.25
13	30	30	0.6	0.5	**0.23**	29	30	29	0.62	0.50	0.24	30	30	30	0.60	0.49	0.23	28	30	30	0.63	0.46	0.25
14	30	60	0.3	0.0	**0.23**	29	29	58	0.33	0.02	0.24	29	29	58	0.33	0.03	0.23	29	29	57	0.34	0.05	0.23
15	30	60	0.3	0.5	**0.23**	30	30	59	0.31	0.50	0.23	30	30	59	0.31	0.50	0.23	30	30	60	0.31	0.50	0.23
16	30	30	0.9	0.0	**0.31**	29	30	31	0.90	0.01	0.32	30	29	29	0.90	0.04	0.31	29	30	31	0.90	0.01	0.32
17	30	30	0.9	0.5	**0.31**	30	30	30	0.90	0.50	0.31	30	30	30	0.90	0.49	0.31	30	30	30	0.90	0.50	0.31
18	30	45	0.6	0.0	**0.31**	30	30	45	0.60	0.01	0.31	29	29	44	0.61	0.02	0.31	29	29	43	0.62	0.04	0.31
19	30	45	0.6	0.5	**0.31**	30	30	45	0.61	0.51	0.31	30	30	45	0.60	0.50	0.31	30	30	45	0.60	0.50	0.31
20	30	60	0.6	0.0	**0.38**	29	29	58	0.62	0.02	0.38	30	29	59	0.61	0.02	0.37	29	29	57	0.63	0.04	0.38
21	30	60	0.6	0.5	**0.38**	30	30	59	0.61	0.50	0.38	30	30	60	0.60	0.50	0.38	30	30	59	0.61	0.50	0.37
22	30	45	0.9	0.0	**0.40**	30	29	44	0.91	0.05	0.40	30	28	43	0.91	0.06	0.40	29	29	44	0.90	0.02	0.40
23	30	45	0.9	0.5	**0.40**	30	30	45	0.90	0.50	0.40	30	30	45	0.90	0.50	0.40	30	30	45	0.90	0.50	0.40
24	30	60	0.9	0.0	**0.47**	30	29	58	0.90	0.02	0.47	30	29	57	0.91	0.05	0.46	29	28	57	0.91	0.06	0.47
25	30	60	0.9	0.5	**0.47**	30	30	59	0.90	0.49	0.47	30	30	60	0.90	0.50	0.47	29	29	58	0.90	0.48	0.47

Table 3B.

Average α, δ, μ, ε_b, ε_s and PIN estimated from 1,000 series of 60 B,S pairs randomly generated from the data generation process assumed by the PIN model with parameters α, δ, μ, $\varepsilon=\varepsilon_b=\varepsilon_s$, if Yan and Zhang 2012 initial values for parameters α, δ, μ, ε_b, ε_s are used. The parameter α is a probability that there is an information event on a given day, δ is a probability that the news is positive, μ is expected number of informed trades on a day with information event, ε_b is expected number of uninformed buys on any day, and ε_s is expected number of uninformed sells on any day. Expected number of uninformed sells is assumed to be the same as the expected number of uninformed buys in the data generation. Parameters are chosen such that PIN implied by the DGP ranges from 0 to 0.47. The factorization used is EHO 2002, EHO 2010, and LK. (The models are sorted by implied PIN, μ, and δ.) "isb" is the percentage of boundary solutions (the solutions in which α or δ equal zero or one).

	parameters of DGP					EHO 2002							EHO 2010							LK						
model	ε	μ	α	δ	PIN	ε_b	ε_s	μ	α	δ	PIN	isb	ε_b	ε_s	μ	α	δ	PIN	isb	ε_b	ε_s	μ	α	δ	PIN	isb
1	30	0	0.0	0.0	0.00	27	27	100	0.45	0.36	0.10	0.68	27	27	100	0.45	0.36	0.10	0.67	28	28	110	0.41	0.35	0.06	0.74
2	30	15	0.3	0.0	0.07	30	26	19	0.41	0.18	0.14	0.65	30	26	19	0.41	0.18	0.14	0.65	30	29	16	0.34	0.07	0.08	0.77
3	30	15	0.3	0.5	0.07	30	30	15	0.32	0.50	0.07	0.02	30	30	15	0.32	0.50	0.07	0.02	30	30	15	0.32	0.50	0.07	0.03
4	30	15	0.6	0.0	0.13	29	24	24	0.63	0.19	0.23	0.59	29	24	24	0.63	0.19	0.23	0.59	30	30	16	0.59	0.04	0.13	0.79
5	30	15	0.6	0.5	0.13	30	30	15	0.61	0.49	0.13	0.00	30	30	15	0.61	0.49	0.13	0.00	30	30	15	0.61	0.49	0.13	0.00
6	30	30	0.3	0.0	0.13	30	25	34	0.39	0.14	0.21	0.81	30	25	33	0.39	0.14	0.20	0.81	30	29	31	0.31	0.02	0.14	0.95
7	30	30	0.3	0.5	0.13	30	30	30	0.30	0.50	0.13	0.00	30	30	30	0.30	0.50	0.13	0.00	30	30	30	0.30	0.50	0.13	0.00
8	30	15	0.9	0.0	0.18	31	22	33	0.81	0.27	0.27	0.50	31	22	33	0.81	0.27	0.27	0.50	32	28	31	0.77	0.13	0.17	0.72
9	30	15	0.9	0.5	0.18	30	30	16	0.90	0.50	0.19	0.10	30	30	16	0.90	0.50	0.19	0.10	30	30	15	0.90	0.50	0.18	0.11
10	30	45	0.3	0.0	0.18	30	25	46	0.41	0.12	0.25	0.83	30	25	46	0.41	0.12	0.25	0.83	30	29	46	0.31	0.02	0.19	0.98
11	30	45	0.3	0.5	0.18	30	30	45	0.30	0.50	0.18	0.00	30	30	45	0.30	0.50	0.18	0.00	30	30	45	0.30	0.50	0.18	0.00
12	30	30	0.6	0.0	0.23	30	29	31	0.60	0.02	0.24	0.94	30	29	31	0.60	0.02	0.24	0.94	30	30	30	0.60	0.01	0.23	0.96
13	30	30	0.6	0.5	0.23	30	30	30	0.60	0.50	0.23	0.00	30	30	30	0.60	0.50	0.23	0.00	30	30	30	0.60	0.50	0.23	0.00
14	30	60	0.3	0.0	0.23	30	29	60	0.33	0.03	0.24	0.96	30	29	60	0.32	0.03	0.24	0.96	30	30	60	0.30	0.01	0.23	0.99
15	30	60	0.3	0.5	0.23	30	30	60	0.30	0.50	0.23	0.00	30	30	60	0.30	0.50	0.23	0.00	30	30	60	0.30	0.50	0.23	0.00
16	30	30	0.9	0.0	0.31	30	24	37	0.89	0.12	0.37	0.75	30	24	37	0.89	0.12	0.38	0.76	31	29	31	0.88	0.03	0.31	0.94
17	30	30	0.9	0.5	0.31	30	30	30	0.90	0.50	0.31	0.01	30	30	30	0.90	0.50	0.31	0.01	30	30	30	0.90	0.50	0.31	0.01
18	30	45	0.6	0.0	0.31	30	28	47	0.61	0.05	0.33	0.92	30	28	47	0.61	0.05	0.33	0.91	30	30	46	0.59	0.01	0.31	0.99
19	30	45	0.6	0.5	0.31	30	30	45	0.60	0.50	0.31	0.00	30	30	45	0.60	0.50	0.31	0.00	30	30	45	0.60	0.50	0.31	0.00
20	30	60	0.6	0.0	0.38	30	29	60	0.60	0.02	0.38	0.98	30	29	60	0.60	0.02	0.38	0.98	30	30	60	0.60	0.01	0.37	0.99
21	30	60	0.6	0.5	0.38	30	30	60	0.60	0.50	0.37	0.00	30	30	60	0.60	0.50	0.37	0.00	30	30	60	0.60	0.50	0.37	0.00
22	30	45	0.9	0.0	0.40	31	30	45	0.89	0.01	0.40	0.98	31	30	45	0.89	0.01	0.40	0.97	31	30	45	0.89	0.01	0.40	0.98
23	30	45	0.9	0.5	0.40	30	30	45	0.90	0.50	0.40	0.00	30	30	45	0.90	0.50	0.40	0.00	30	30	45	0.90	0.50	0.40	0.00
24	30	60	0.9	0.0	0.47	31	30	60	0.89	0.01	0.47	0.98	31	30	60	0.89	0.01	0.47	0.98	31	30	60	0.89	0.01	0.47	0.98
25	30	60	0.9	0.5	0.47	30	30	60	0.90	0.50	0.47	0.00	30	30	60	0.90	0.50	0.47	0.00	30	30	60	0.90	0.50	0.47	0.00

Table 4.:

Average PIN estimated from 1,000 series of 60 B,S pairs randomly generated from the data generation process assumed by the PIN model with parameters α, δ, μ, ε. The parameter α is a probability that there is an information event on a given day, δ is a probability that the news is positive, μ is expected number of informed trades on a day with information event, ε is expected number of uninformed buys on any day. Expected number of uninformed sells is assumed to be the same as the expected number of uninformed buys. Parameters are chosen such that PIN implied by the DGP ranges from 0 to 0.47. The table shows three different choices of initial values to estimation: values supporting zero PIN, randomized values, and values supporting high PIN. For each case of initial parameters, three factorizations of the likelihood functions are used: EHO 2002, EHO 2010, and LK. The SAS code for each factorization is provided in Appendix B. (The models are sorted by implied PIN, μ, and δ.)

model		parameters of DGP				null hypothesis initial values: α=0.5, δ=0.5, μ=0, ε_b=average B, ε_s=average S (assumes PIN=0)			random initial values: α, δ is random number 0 to 1; ε_b, ε_s, μ is random integer 0 to global average of (B + S)			high-PIN initial values: α=1, δ=0.5, μ=average(B+S)/2 ε_b=average B/2, ε_s=average S/2 (assumes PIN=0.5)		
	ε	μ	α	δ	implied PIN	PIN EHO2002	PIN EHO2010	PIN LK	PIN EHO2002	PIN EHO2010	PIN LK	PIN EHO2002	PIN EHO2010	PIN LK
1	30	0	0.0	0.0	**0.00**	0.00	0.00	0.00	*0.11*	*0.11*	*0.11*	*0.32*	*0.32*	*0.32*
2	30	15	0.3	0.0	**0.07**	0.00	0.00	0.00	*0.12*	*0.12*	*0.12*	*0.17*	*0.17*	*0.17*
3	30	15	0.3	0.5	**0.07**	0.00	0.00	0.00	*0.11*	*0.11*	*0.11*	*0.17*	*0.17*	*0.17*
4	30	15	0.6	0.0	**0.13**	0.00	0.00	0.00	0.15	0.15	0.15	*0.19*	*0.19*	*0.19*
5	30	15	0.6	0.5	**0.13**	0.00	0.00	0.00	0.14	0.14	0.14	0.15	0.15	0.15
6	30	30	0.3	0.0	**0.13**	0.00	0.00	0.00	*0.16*	*0.16*	*0.16*	0.13	0.13	0.13
7	30	30	0.3	0.5	**0.13**	0.00	0.00	0.00	*0.16*	*0.16*	*0.17*	0.13	0.13	0.13
8	30	15	0.9	0.0	**0.18**	0.00	0.00	0.00	0.17	0.17	0.18	*0.27*	*0.27*	*0.27*
9	30	15	0.9	0.5	**0.18**	0.00	0.00	0.00	0.18	0.18	0.18	0.19	0.19	0.19
10	30	45	0.3	0.0	**0.18**	0.00	0.00	0.00	0.20	0.20	*0.21*	0.18	0.18	0.18
11	30	45	0.3	0.5	**0.18**	0.00	0.00	0.00	0.20	0.20	0.20	0.19	0.19	0.19
12	30	30	0.6	0.0	**0.23**	0.00	0.00	0.00	0.22	0.22	0.22	0.23	0.23	0.23
13	30	30	0.6	0.5	**0.23**	0.00	0.00	0.00	0.25	0.25	0.25	0.23	0.23	0.23
14	30	60	0.3	0.0	**0.23**	0.00	0.00	0.00	0.24	0.24	0.24	0.23	0.23	0.23
15	30	60	0.3	0.5	**0.23**	0.00	0.00	0.00	0.24	0.24	0.24	0.23	0.23	0.23
16	30	30	0.9	0.0	**0.31**	0.00	0.00	0.00	*0.23*	*0.23*	*0.23*	0.31	0.31	0.31
17	30	30	0.9	0.5	**0.31**	0.00	0.00	0.00	0.31	0.31	0.31	0.31	0.31	0.31
18	30	45	0.6	0.0	**0.31**	0.00	0.00	0.00	0.30	0.30	0.30	0.31	0.31	0.31
19	30	45	0.6	0.5	**0.31**	0.00	0.00	0.00	0.31	0.31	0.31	0.31	0.31	0.31
20	30	60	0.6	0.0	**0.38**	0.00	0.00	0.00	0.37	0.37	0.37	0.37	0.37	0.37
21	30	60	0.6	0.5	**0.38**	0.00	0.00	0.00	0.38	0.38	0.38	0.37	0.37	0.37
22	30	45	0.9	0.0	**0.40**	0.00	0.00	0.00	*0.33*	*0.33*	*0.33*	0.40	0.40	0.40
23	30	45	0.9	0.5	**0.40**	0.00	0.00	0.00	0.40	0.40	0.40	0.41	0.41	0.41
24	30	60	0.9	0.0	**0.47**	0.00	0.00	0.00	*0.43*	*0.43*	*0.43*	0.47	0.47	0.47
25	30	60	0.9	0.5	**0.47**	0.00	0.00	0.00	0.47	0.47	0.47	0.47	0.47	0.47

Table 5.:

Standard deviation of PIN estimated from 1,000 series of 60 B,S pairs randomly generated from the data generation process assumed by the PIN model with parameters α, δ, μ, ε. The parameter α is a probability that there is an information event on a given day, δ is a probability that the news is positive, μ is expected number of informed trades on a day with information event, ε is expected number of uninformed buys on any day. Expected number of uninformed sells is assumed to be the same as the expected number of uninformed buys. Parameters are chosen such that PIN implied by the DGP ranges from 0 to 0.47. The table shows six different choices of initial values to estimation: A. values equal to DGP parameters, B. no initial values, C. initial values as in Yan and Zhang 2012, D. values supporting zero PIN, E. randomized values, and F. values supporting high PIN. For each case of initial parameters, three factorizations of the likelihood functions are used: 1. EHO 2002, 2. EHO 2010, and 3. LK. The SAS code for each factorization is provided in Appendix B. (The models are sorted by implied PIN, μ, and δ.)

model	e	m	a	d	PIN	A PIN1	PIN2	PIN3	B PIN1	PIN2	PIN3	C PIN1	PIN2	PIN3	D PIN1	PIN2	PIN3	E PIN1	PIN2	PIN3	F PIN1	PIN2	PIN3
1	30	0	0	0.0	0.00	0.00	0.00	0.00	0.06	0.08	0.06	0.17	0.17	0.08	0.00	0.00	0.00	0.18	0.18	0.17	0.09	0.10	0.09
2	30	15	0.3	0.0	0.07	0.02	0.02	0.02	0.12	0.10	0.09	0.16	0.16	0.06	0.00	0.00	0.00	0.17	0.17	0.17	0.13	0.13	0.13
3	30	15	0.3	0.5	0.07	0.02	0.02	0.02	0.10	0.10	0.10	0.04	0.04	0.02	0.00	0.00	0.00	0.16	0.16	0.15	0.13	0.13	0.13
4	30	15	0.6	0.0	0.13	0.03	0.03	0.03	0.09	0.10	0.10	0.21	0.21	0.06	0.00	0.00	0.00	0.18	0.18	0.17	0.10	0.10	0.10
5	30	15	0.6	0.5	0.13	0.02	0.02	0.02	0.05	0.05	0.05	0.02	0.02	0.02	0.00	0.00	0.00	0.15	0.15	0.15	0.06	0.06	0.06
6	30	30	0.3	0.0	0.13	0.02	0.02	0.02	0.05	0.06	0.06	0.18	0.18	0.08	0.00	0.00	0.00	0.12	0.12	0.12	0.03	0.03	0.03
7	30	30	0.3	0.5	0.13	0.02	0.02	0.02	0.04	0.04	0.05	0.02	0.02	0.02	0.00	0.00	0.00	0.11	0.11	0.11	0.03	0.03	0.03
8	30	15	0.9	0.0	0.18	0.06	0.06	0.06	0.11	0.09	0.11	0.22	0.22	0.12	0.00	0.00	0.00	0.20	0.20	0.20	0.10	0.10	0.10
9	30	15	0.9	0.5	0.18	0.02	0.02	0.02	0.05	0.06	0.05	0.05	0.05	0.02	0.00	0.00	0.00	0.15	0.15	0.15	0.04	0.04	0.04
10	30	45	0.3	0.0	0.18	0.03	0.03	0.03	0.05	0.06	0.05	0.16	0.16	0.08	0.00	0.00	0.00	0.09	0.09	0.09	0.03	0.03	0.03
11	30	45	0.3	0.5	0.18	0.03	0.03	0.03	0.05	0.03	0.05	0.03	0.03	0.03	0.00	0.00	0.00	0.07	0.07	0.08	0.03	0.03	0.03
12	30	30	0.6	0.0	0.23	0.02	0.02	0.02	0.07	0.04	0.07	0.08	0.08	0.04	0.00	0.00	0.00	0.14	0.14	0.13	0.02	0.02	0.02
13	30	30	0.6	0.5	0.23	0.02	0.02	0.02	0.03	0.03	0.06	0.02	0.02	0.02	0.00	0.00	0.00	0.08	0.08	0.07	0.02	0.02	0.02
14	30	60	0.3	0.0	0.23	0.04	0.04	0.04	0.06	0.05	0.05	0.08	0.08	0.05	0.00	0.00	0.00	0.08	0.08	0.07	0.04	0.04	0.04
15	30	60	0.3	0.5	0.23	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.06	0.06	0.06	0.04	0.04	0.04
16	30	30	0.9	0.0	0.31	0.04	0.04	0.04	0.06	0.05	0.07	0.16	0.16	0.07	0.00	0.00	0.00	0.18	0.18	0.18	0.04	0.04	0.04
17	30	30	0.9	0.5	0.31	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.07	0.07	0.07	0.02	0.02	0.02
18	30	45	0.6	0.0	0.31	0.03	0.03	0.03	0.03	0.04	0.03	0.10	0.10	0.04	0.00	0.00	0.00	0.08	0.08	0.08	0.03	0.03	0.03
19	30	45	0.6	0.5	0.31	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.00	0.00	0.00	0.04	0.04	0.04	0.03	0.03	0.03
20	30	60	0.6	0.0	0.38	0.03	0.03	0.03	0.03	0.03	0.04	0.05	0.05	0.04	0.00	0.00	0.00	0.05	0.05	0.05	0.03	0.03	0.03
21	30	60	0.6	0.5	0.38	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.00	0.00	0.00	0.03	0.03	0.03	0.03	0.03	0.03
22	30	45	0.9	0.0	0.40	0.03	0.03	0.03	0.05	0.05	0.04	0.07	0.07	0.06	0.00	0.00	0.00	0.17	0.17	0.16	0.04	0.04	0.04
23	30	45	0.9	0.5	0.40	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.03	0.03	0.03	0.02	0.02	0.02
24	30	60	0.9	0.0	0.47	0.02	0.02	0.02	0.04	0.06	0.06	0.06	0.06	0.06	0.00	0.00	0.00	0.13	0.13	0.14	0.04	0.04	0.04
25	30	60	0.9	0.5	0.47	0.02	0.02	0.02	0.11	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02

Appendix A.: The VBA code used to generate random B,S series.

```vba
Application.ScreenUpdating = False
Application.DisplayAlerts = False
MkDir "C:\Users\user\Desktop\PINsimulation"

n = 2 'set desired number of random series to be generated
numbers = 60 'set the length of the series: number of days for which daily number of buys
and sells is to be generated
epsilon = 30 'average number of uninformed buys and sells (assumed equal)

'first, generate B and S without informed trades;
    For I = 1 To n
        Range("A" & (I - 1) * numbers + 1 & ":" & "A" & I * numbers) = I
        Application.Run "ATPVBAEN.XLAM!Random", Range("B" & (I - 1) * numbers + 1), 1,
numbers, 5, , epsilon
        Application.Run "ATPVBAEN.XLAM!Random", Range("C" & (I - 1) * numbers + 1), 1,
numbers, 5, , epsilon
    Next I

    Columns("A:C").Select
    Selection.Copy
    Workbooks.Add
    ActiveSheet.Paste
    ActiveWorkbook.SaveAs Filename:="C:\Users\user\Desktop\PINsimulation\BS" & "_epsilon"
& epsilon & "_alpha0%_delta0%_mu0_0.csv", FileFormat:=xlCSV
    ActiveWindow.Close

    'next, add informed trades to B or S according to alpha, delta, mu
    For d = 1 To 2
    Delta = d / 2 - 0.5 'delta is either zero or 0.5
        For a = 3 To 9 Step 3
        alpha = a / 10
            For mu = 15 To 60 Step 15
                For I = 1 To n
                    Application.Run "ATPVBAEN.XLAM!Random", Range("D" & (I - 1) * numbers +
1), 1, numbers, 5, , mu
                Next I
                For j = 1 To n * numbers
                    If Rnd() < alpha Then
                        If Rnd() > Delta Then
                        Range("E" & j) = Range("B" & j) + Range("D" & j) 'signal is
positive: add "mu" to buys
                        Range("F" & j) = Range("C" & j)
                        Else
                        Range("E" & j) = Range("B" & j)
                        Range("F" & j) = Range("C" & j) + Range("D" & j) 'signal is
negative: add "mu" to sells
                        End If
                    Else
                        Range("E" & j) = Range("B" & j)
                        Range("F" & j) = Range("C" & j)
                    End If
                    Range("D" & j) = Range("A" & j)
                 Next j

                Columns("D:F").Select
                Selection.Cut
                Workbooks.Add
                ActiveSheet.Paste
                ChDir "C:\Users\user\Desktop"

                ActiveWorkbook.SaveAs Filename:="C:\Users\user\Desktop\PINsimulation\BS" &
"_epsilon" & epsilon & "_alpha" & alpha * 100 & "%_delta" & Delta * 100 & "%_mu" & mu &
"_1.csv", FileFormat:=xlCSV
                ActiveWindow.Close
                Range("A1").Select
            Next mu
        Next a
    Next d

    Application.ScreenUpdating = True
    Application.DisplayAlerts = True

End Sub
```

Appendix B.: The SAS code to implement particular likelihood function factorizations used in this study.

1. EHO 2002 factorization:

```
proc nlp data=mydata&i tech=newrap maxstep=5 maxiter=100 maxfunc=250
noprint out=pinest&i (keep = fname id a d e f m loglik) INEST=initial;
 by fname id;
 max loglik;
 decvar a d e f m;
 x1 = e / (m + e);
 x2 = f / (m + f);
 bounds   0 < e f, 0<= m, 0 <= a d <= 1;

 loglik = -e - f + MM * (log(x1)+log(x2)) + B * log(m + e) + S * log(m + f) +
          log( a * (1 - d) * exp(-m) * (x2 ** (S - MM ))*(x1**(-MM)) +
          a * d * exp(-m) * (x1 ** (B - MM))*(x2**(-MM))  +
          (1 - a) * (x2 ** (S - MM))*(x1**(B-MM)));
run;
```

2. EHO 2010 factorization:

```
proc nlp data=mydata&i tech=newrap maxstep=5 maxiter=100 maxfunc=250
noprint out=pinest&i (keep = fname id a d e f m loglik) INEST=initial;
 by fname id;
 max loglik;
 decvar a d e f m;
 MM = min(B,S) + max(B,S) / 2;
 X_B = e / (m + e);
 X_S = f / (m + f);
 negative_MM = -1 * MM;
 negative_m = -1 * m;
 one_minus_d = 1 - d;
 one_minus_a = 1 - a;
 S_minus_MM= S-MM;
 B_minus_MM= B-MM;
 bounds   0 < e f, 0<= m, 0 <= a d <= 1;

loglik = -e + MM * log(X_B) + B*log(m+e)
         -f + MM * log(X_S) + S*log(m+f)
         + log(a*one_minus_d*X_S**S_minus_MM*X_B**negative_MM*exp(negative_m)
           + a*d*X_B**B_minus_MM*X_S**negative_MM*exp(negative_m)
           + one_minus_a* X_B**B_minus_MM* X_S**S_minus_MM);
run;
```

3. LK factorization:

```
proc nlp data=mydata&i tech=newrap maxstep=5 maxiter=100 maxfunc=250
noprint out=pinest&i (keep = fname id a d e f m loglik) INEST=initial;
 by fname id;
 max loglik;
 decvar a d e f m;
     one_minus_d = 1 - d;
 one_minus_a = 1 - a;
 e1 = -1*m - B*log(1+m/e);
 e2 = -1*m - S*log(1+m/f);
 e3 = -B*log(1+m/e) - S*log(1+m/f);
 emax = max(e1, e2, e3);
 bounds   0 < e f, 0<= m, 0 <= a d <= 1;

loglik = -e + B*log(m+e)
         -f + S*log(m+f)
         + emax
         + log(a * one_minus_d * exp(e2 - emax)
           + a * d * exp(e1 - emax)
           + one_minus_a * exp(e3 - emax));
run;
```