Journal of Mathematical Finance, 2019, 9, 637-666
https://www.scirp.org/journal/jmf
ISSN Online: 2162-2442
ISSN Print: 2162-2434

A New Way to Compute the Probability of Informed Trading

Antoine Bambade[1,2]

[1]Lawrence Berkeley National Laboratory, SDM Group, California, USA
[2]Department of Applied Mathematics, Ecole Polytechnique, Palaiseau, France
Email: antoine.bambade@polytechnique.edu

How to cite this paper: Bambade, A. (2019) A New Way to Compute the Probability of Informed Trading. *Journal of Mathematical Finance*, **9**, 637-666.
https://doi.org/10.4236/jmf.2019.94032

Received: July 10, 2019
Accepted: October 22, 2019
Published: October 25, 2019

Abstract

Volume-Synchronized Probability of Informed Trading (VPIN) is a tool designed to predict extreme events like flash crashes in high-frequency trading. Its aim is to estimate the Probability of Informed Trading (PIN), which was built from a probabilistic framework. Some concerns have been raised about its theoretical foundations and its reliability. More precisely, it has been shown that theoretically the VPIN does not approximate the PIN as the PIN has been built with a time-clock framework and the VPIN with a volume clock one. On a practical point of view, the VPIN has been found to be sensitive to the starting point of computation of a data set and to different parameters, such as the classification rule. In this paper, in order to improve the PIN theoretical framework, we firstly analyze the theoretical foundations of the PIN and the VPIN models to have a better view of all its different assumption subtleties. It secondly makes it possible to point out some approximation flaws in the formula used to approximate the PIN and to propose another exact way to compute the PIN. All different results are illustrated with simulations.

Keywords

High-Frequency Data, Probability of Informed Trading, Flash Crash, Flow Toxicity, Volume Imbalance

1. Introduction

The amount of trading data has exploded in finance thanks to the continuing progress of high frequency techniques. It constrains practitioners to use more and more state-of-the-art algorithms to deal with this overwhelming amount of information. Computers and algorithms are more and more efficient, but still decision making is based on both the quantity and the quality of information.

Thus, errors and speculations that can make the financial market toxic, *i.e.* conducive to crashes, are still possible. Examples in the past, such as the "Flash Crash" of May 6, 2010, have shown that algorithmic trading in finance has made it possible to introduce new kind of crashes characterized by their suddenness. Such quick crashes seem dangerous because of a kind of inherent unpredictability. However, theoretical framework to model this new phenomenon exists.

Easley, Engle, O'Hara and Wu [1] designed a model of the high-frequency financial market based on flows of informed and uninformed traders. In this model, informed traders are aware of the evolution of the price in the future and thus of which decision takes (buy or sell). The authors managed to show that information is a key parameter of the spread between ask and bid of prices, as they demonstrate that the probability of being informed within their theoretical framework is proportionally linked with it. They named this key parameter the Probability of Informed Trading (PIN). A high value of the PIN is an indicator of the level of toxicity of this high frequency trading market, as it would mean it relies on too many informed traders. Later, Easley, Lopez de Prado, O'Hara [2] [3] designed a tool, nicknamed Volume-synchronized Probability of Informed Trading (VPIN), supposed to approximate the PIN. It appeared it could predict the "Flash Crash" of May, 6 2010 a few hours before it happened [4]. A number of papers have been written [5] [6] [7], and it is proposed to use it for regulation through a VPIN contract [4] [8]. However, critics pointed out some flaws, questioning its reliability. For example, Andersen and Bondarenko have shown [9] that the VPIN is quite sensitive to the starting point of when one starts computing the VPIN on a data set. It indeed questions the VPIN prediction quality. Moreover, they have also shown that the VPIN is sensitive to other parameters, such as the trade classification rule used [10], or how one defines the average daily volume of trades [11]. Changing the classification rule may drastically change the VPIN behavior [12]. Tomas Pöppe, Sebastian Moos and Dirk Schiereck have arrived to the same conclusions with a different approach. Using a different classification rule can change the VPIN prediction power toward a crash (in their paper a German blue-chip stock) [13]. Besides, controlling ex-ante parameters seem to give poorer prediction quality [10] [11]. This point has also been checked by D. Abad, M. Massot and R. Pascual [12]. Controlling for ex-ante realized volatility, and trading intensity, as did T. G. Andersen and O. Bondarenko [11], prediction quality seems to vanish. More deeper, they have also underlined that it is not obvious how one should define a VPIN prediction, analyzing more precisely toxic and non-toxic halts, as well as toxic events. Furthermore, Torben G. Andersen and Oleg Bondarenko interpret the VPIN as being too sensitive to trading intensity. They have also explained the VPIN metric is sometimes unexpectedly correlated with other usual ones (such as VIX or RV) [9] [10]. Moreover, it has been shown [14] [15] that the VPIN does not approximate the PIN, as the PIN was built on a time-clock theoretical framework, and the VPIN with a volume-clock paradigm. In this study, we propose another way to estimate the PIN within its original time-clock framework.

The purpose of this paper is to improve the PIN theoretical framework. Some concerns have been raised about its theoretical foundations. For this reason we assess step by step all the different theoretical ideas of the PIN model. More precisely, we firstly want to explicit all the theoretical framework of the PIN and the VPIN model to have a better view of all its different assumption subtleties. It secondly makes it possible to point out some approximation errors in the formula used to approximate the PIN and to propose another exact way to compute the PIN. In the following, we first recall the PIN model (Section 2). Second, after introducing the VPIN original ideas we analyze the original first order approximation and then recall the difference of time clock and volume clock paradigm (Section 3). Finally, we suggest another way to compute the PIN (Section 4).

2. The PIN Model

2.1. The Time-Clock Framework

The Probability of Informed Trading (PIN) is computed on a simple model of information among traders [16]. Let's describe it with the following tree below (**Figure 1**), originally designed in [16]. Suppose prior to the beginning of any trading day, Nature determines whether an information event is relevant to the value of the asset to occur. Suppose information events are independently distributed and occur with a Bernoulli probability of value α, which can be seen on the first two branches on the left-hand side of the tree. These events are good news with a Bernoulli probability $1-\delta$ (*i.e.* signal High), or bad news with probability δ (*i.e.* signal Low). After the end of trading on any day, and before Nature moves again, the full information value of the asset is realized. Hence, for any of the three leaves of the tree in **Figure 1**, an informed trader would know which action to take. Trade arises from both informed traders (those who have seen any signal) and uninformed traders. On any day, arrivals of uninformed buyers and uninformed sellers are described by independent Poisson processes of respective intensity ϵ and μ. Individuals trade a single risky asset and money with a market maker over $i=1,\cdots,I$ trading days. Within any trading day, time is continuous and it is indexed by $t\in[0,T]$. Let's define for $t\in[0,T]$, for a given trading day, S_t and B_t the events that an order of respectively a sell and a buy arrive at time *t*. Let $P_t=\left(P_t(n),P_t(b),P_t(g)\right)$ be the market maker's prior belief about the events "no news" (*n*) "bad news" (*b*) and "good news" (*g*) at time *t*[1]. Within this model we compute the spread at *t* Σ_t which is equal to a_t-b_t, where a_t and b_t are the ask and bid at time *t* (respectively the minimum price a seller is willing to receive and the maximum price a buyer

[1] We summarize here the theoretic framework as described in [16]. Formally, considering the random variables corresponding to order arrival of sells and buys St and Bt we associate the canonical respective filtrations to define later conditioned expectations. They are still noted as the events "St" and "Bt".

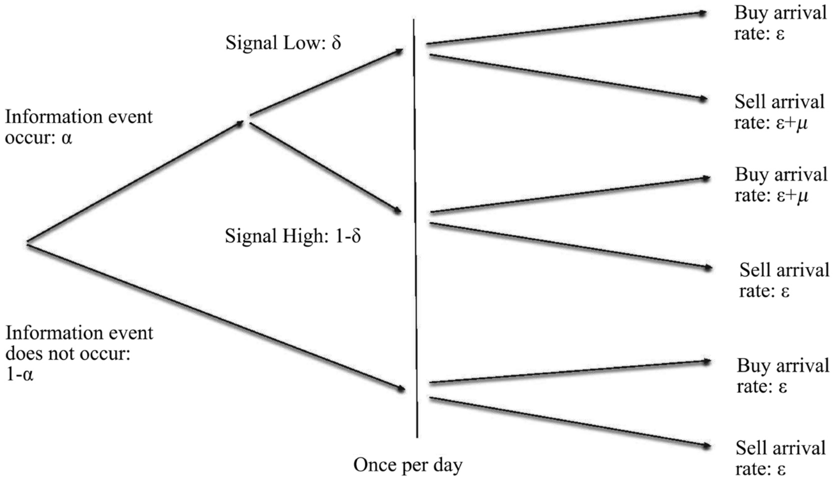

Figure 1. A tree summarizing the theoretical framework.

is willing to pay). Within this framework b_t is the expectation of the asset value, we denote V_t, conditional on the history prior to t and on sell order S_t. Similarly, a_t is the expectation of V_t conditional on the history prior to t and on buy order B_t. Let note $\bar{V}$, V^* and $\underline{V}$ respectively the value of the asset under the conditions of good new, no information and bad new. We have of course the following inequalities: $\underline{V} \le V^* \le \bar{V}$.

2.2. Computation of the Spread

We explicit now more the content of [3]. Let's compute the bid, the ask follows exactly the same idea[2]:

$$b_t = E\left(V_t \mid t, S_t\right).$$

It can be re-written this way using the different possibilities of the tree on an event:

$$\begin{aligned} b_t &= E\left(V_t \mid t, S_t, n\right) P_t\left(n \mid S_t\right) + E\left(V_t \mid t, S_t, g\right) P_t\left(g \mid S_t\right) + E\left(V_t \mid t, S_t, b\right) P_t\left(b \mid S_t\right) \\ &= V^* P_t\left(n \mid S_t\right) + \bar{V} P_t\left(g \mid S_t\right) + \underline{V} P_t\left(b \mid S_t\right). \end{aligned}$$

Let's compute the first term $P_t\left(n \mid S_t\right)$, others follow the same idea. Using Bayes rule one finds the following:

$$P_t\left(n \mid S_t\right) = \frac{P_t\left(S_t \mid n\right) P_t(n)}{P_t\left(S_t\right)},$$

so, by decomposing the denominator:

$$P_t\left(n \mid S_t\right) = \frac{P_t\left(S_t \mid n\right) P_t(n)}{P_t\left(S_t \mid n\right) P_t(n) + P_t\left(S_t \mid g\right) P_t(g) + P_t\left(S_t \mid b\right) P_t(b)}.$$

Let's have a look at the term $P_t\left(S_t \mid n\right)$ which is the probability at t that there will be a sell order at t under the constraints of no news. $P_t\left(S_t \mid n\right)$ is a transition rate. To compute it, one must first calculate the transition probability for a

[2] We use the same notations as the author, distinguishing the events "t" and "S_t".

strictly positive time length let say *h*. Formally, if one notes N_t the number of jumps of the corresponding Poisson process up to t under conditions of no events, we know its intensity is ϵt under the constraint of no news. For any *h* strictly positive and small enough we look to the limit of the number $\frac{P(N_t - N_{t-h} \geq 1 | n)}{h}$ when *h* goes to zero remaining strictly positive, which defines the transition rate. At first order on *h*, one finds:

$$P(N_t - N_{t-h} \geq 1 | n) = 1 - \mathrm{e}^{\epsilon h} = \epsilon h + o(h).$$

Dividing by *h*, one re-finds indeed the intensity of the Poisson process, which is a special case of a Markov jump process. Applying the same for other cases ("bad event", "good event"), we have finally the following:

$$\begin{aligned} &P_t(S_t | n) P_t(n) + P_t(S_t | g) P_t(g) + P_t(S_t | b) P_t(b) \\ &= P_t(n)\epsilon + P_t(g)\epsilon + P_t(b)(\mu + \epsilon). \end{aligned}$$

As the probabilities with ϵ sum to one we get the following expression:

$$P_t(S_t | n) P_t(n) + P_t(S_t | g) P_t(g) + P_t(S_t | b) P_t(b) = \epsilon + P_t(b)\mu.$$

Finally the bid has this expression:

$$b_t = \frac{P_t(n)\epsilon V^* + P_t(b)(\epsilon + \mu)\underline{V} + P_t(g)\epsilon \bar{V}}{\epsilon + P_t(b)\mu}.$$

With the same reasoning the ask has this expression:

$$a_t = \frac{P_t(n)\epsilon V^* + P_t(b)\epsilon \underline{V} + P_t(g)(\epsilon + \mu)\bar{V}}{\epsilon + P_t(g)\mu}.$$

Actually one may simplify a bit these expressions as the expectation of V has the following form:

$$E(V_t) = V^* P_t(n) + \underline{V} P_t(b) + \bar{V} P_t(g).$$

We find:

$$b_t = \frac{\mu \underline{V} P_t(b)}{\epsilon + P_t(b)\mu} + \frac{\epsilon}{\epsilon + P_t(b)\mu} E(V_t),$$

and:

$$a_t = \frac{\mu \bar{V} P_t(g)}{\epsilon + P_t(g)\mu} + \frac{\epsilon}{\epsilon + P_t(g)\mu} E(V_t).$$

So the spread equals to:

$$\Sigma_t = a_t - b_t = E(V_t)\left(\frac{\epsilon}{\epsilon + P_t(g)\mu} - \frac{\epsilon}{\epsilon + P_t(b)\mu}\right) + \frac{\mu \bar{V} P_t(g)}{\epsilon + P_t(g)\mu} - \frac{\mu \underline{V} P_t(b)}{\epsilon + P_t(b)\mu}.$$

In the special case where $P_t(g) = P_t(b)$ one finds the following simple form:

$$\Sigma_t = \frac{\mu P_t(g)}{\epsilon + P_t(g)\mu}(\bar{V} - \underline{V}) = \frac{\mu(1 - P_t(n))}{2\epsilon + (1 - P_t(n))\mu}(\bar{V} - \underline{V}).$$

If we make the hypothesis that $P_t(n) = P_0(n) = 1 - \alpha$ is constant, then we have the following:

$$\Sigma_t = \frac{\mu\alpha}{2\epsilon + \alpha\mu}\left(\bar{V} - \underline{V}\right) = \text{thePIN}\left(\bar{V} - \underline{V}\right).$$

Thus, with the assumptions: $P_t(g) = P_t(b) = \delta = 1 - \delta$, *i.e.* $\delta = \frac{1}{2}$ and $P_t(n) = P_0(n) = 1 - \alpha$, the PIN equals the following:

$$\text{PIN} = \frac{\mu\alpha}{2\epsilon + \mu\alpha}.$$

We will keep the same hypothesis for the rest of the paper.

3. Analysis of the First Order Approximate within the Time-Clock Framework

The idea behind the VPIN is to find an easy way to compute the last above expression of the PIN using a volume-clock paradigm. More precisely, it aims at finding a way to easily compute the expressions obtained for the numerator $\alpha\mu$ and denominator ($\alpha\mu + 2\epsilon$). The key heuristic behind the VPIN is to take advantage of a supposedly good property of the expectation of the absolute difference between Poisson random variable within a volume-clock framework to approximate $\alpha\mu$, *i.e.*: $E\left(\left|X - Y\right|\right)$, where X and Y are Poisson variables. We will see this heuristic does not really make it possible to conclude as expected. More precisely, in the first subsection we will see which idea has been used to approximate the PIN within a time-clock framework. Secondly, we will see that first-order approximations used are not correct as the framework does not verify a required hypothesis. We analyze more precisely the first order approximates which can be made in the time-clock framework. In the third subsection, we describe the volume-clock framework and explain why its hypotheses lead to different results compared to the time-clock framework. Finally, we illustrate our results with simulations.

3.1. The Design of a New Heuristic

In the first subsection we see which idea has been used to approximate the PIN within a time-clock framework. We refer now to the related work of Easley *et al.* [1]. Considering the previous framework the probability to obtain on the same time $y_t = (S, B)$, S sells and B buys for day t of length one is:

$$P\left(y_t = (S,B)\right) = (1-\delta)\alpha e^{-(\mu+2\epsilon)}\frac{(\mu+\epsilon)^B \epsilon^S}{B!S!} + (1-\alpha)e^{-2\epsilon}\frac{\epsilon^{B+S}}{B!S!}$$
$$+ \alpha\delta e^{-(\mu+2\epsilon)}\frac{(\mu+\epsilon)^S \epsilon^B}{B!S!}.$$

So, if one notes $TT = S + B$ the total number of trades for this day, one finds, conditioning by all possibilities of the model:

$$E(TT) = \alpha(1-\delta)E(TT \mid g) + \alpha\delta E(TT \mid b) + (1-\alpha)E(TT \mid n).$$

S and B are independant Poisson process, so one can sum in each case their respective intensities to find new Poisson processes. Thus:

$$E\left(TT\right)=\alpha\left(1-\delta\right)\left(\epsilon+\mu+\epsilon\right)+\alpha\delta\left(\mu+\epsilon+\epsilon\right)+\left(1-\alpha\right)\left(\epsilon+\epsilon\right),$$

i.e.

$$E\left(TT\right)=\alpha\mu+2\epsilon.$$

Note the following:

- Remark 1: the time period is fixed, thus S and B can take whatever possible positive integer values, which won't be the case if S + B was fixed.
- Remark 2: intensities are rates, thus the equation has a meaning because one implicitly multiplies it by one (trading day).

The authors propose to compute the expectation of the absolute value of the following random number K = S − B with an approximate. This is the intuition behind the computation of the VPIN. They refer to the following paper of Katti [17] but do not explicit any calculus. They assert that $E\left(\left|K\right|\right)=\alpha\mu$ thanks to a first order approximation without explaining what it does mean. Let's first describe the content of this reference and assumptions assumed. Then let's describe which computations are involved within this time-clock framework.

3.1.1. Katti's Reference Assumptions

The reference proposes several ways to compute the expectation of the absolute value of two random variables that follow same discrete positive distribution but with possibly different parameters. The case of Poisson processes is treated. Let's describe the beginning of Katti's paper [17]. Let's note X_1 and X_2 two Poisson random variables of intensity λ_1 and λ_2. We would like to compute the following number $\Delta_1=E\left|X_1-X_2\right|$. One can write the following:

$$\begin{aligned}\Delta_1&=\sum_{i,k}kP\left(X_2-X_1=k\mid X_1=i\right)P\left(X_1=i\right)\\&\quad+\sum_{i,k}kP\left(X_1-X_2=k\mid X_2=i\right)P\left(X_2=i\right)\\&=\sum_{i,k}kP_i^1P_{i+k}^2+\sum_{i,k}kP_i^2P_{i+k}^1,\end{aligned}$$

where the summations are over $1,2,3,\cdots$ and $P_i^1=\mathrm{e}^{-\lambda_1}\dfrac{\lambda_1^i}{i!}$ and $P_i^2=\mathrm{e}^{-\lambda_2}\dfrac{\lambda_2^i}{i!}$.

Then, one can develop it as follows:

$$\Delta_1=\mathrm{e}^{-\lambda_1-\lambda_2}\left(\sum_{i=0}^{\infty}\sum_{i=0}^{\infty}\frac{k\left(\lambda_1\lambda_2\right)^i\lambda_2^k}{i!\left(i+k\right)!}+\sum_{i=0}^{\infty}\sum_{i=0}^{\infty}\frac{k\left(\lambda_1\lambda_2\right)^i\lambda_1^k}{i!\left(i+k\right)!}\right)=\mathrm{e}^{-\lambda_1-\lambda_2}\left(A_1+B_2\right),$$

with A_1 and B_1 the two different sums. The author, in order to simplify the calculus and use a trick, makes the following assumptions: $\lambda_1\lambda_2=\nu$, where ν is a constant not linked anymore to λ_2 nor λ_1. It implies thus a relation between the two variables (for example $\lambda_1=\dfrac{1}{\lambda_2}$). Thanks to this assumption he can do the following:

$$A_1=\left(\lambda_2\frac{\delta}{\delta\lambda_2}\right)\left(\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\frac{\nu^i\lambda_2^k}{i!(i+k)!}\right)=\left(\lambda_2\frac{\delta}{\delta\lambda_2}\right)A_0,$$

say, with:

$$A_0 = \sum_{i=0}^{\infty} \frac{\nu^i}{(i!)^2} F_1(1; i+1; \lambda_2),$$

where $F_1(\alpha, \gamma, x)$ is a confluent hypergeometric function. Operating by ($\lambda_2 \frac{\delta}{\delta\lambda_2}$) it finally leads to:

$$A_1 = \left(\lambda_2 - \frac{\nu}{\lambda_2}\right) A_0 + \nu e^{2\sqrt{\nu}} F_1\left(\frac{3}{2}; 3; -4\sqrt{\nu}\right) + \frac{\nu}{\lambda_2} F_1\left(\frac{1}{2}; 1; -4\sqrt{\nu}\right).$$

The particular case of $\lambda_1 = \lambda_2 = \lambda$ cannot be treated with this trick because it would imply equal numbers are linked by an inverse relation, so that the product is independant of λ_2. But $\lambda_1 = \lambda_2 = \sqrt{\nu}$, ν is not anymore a constant of the main parameters λ_1 and λ_2, so applying the operator does not give the previous results. One may use here another reference, one cited by Katti [18]. We will detail later the same ideas for our precise the VPIN framework. Anyway, this case leads to the following:

$$\Delta_1 = 2\lambda e^{-2\lambda} \left(I_0(2\lambda) + I_1(2\lambda)\right),$$

where $I_n(x) = \sum_{i=0}^{\infty} \frac{\left(\frac{x}{2}\right)^{n+2i}}{i!(n+i)!}$ is a modified Bessel function of first kind.

3.1.2. How to Use as Far as Possible References' Work to Approximate the VPIN in a Time-Clock Framework

First, let's put ourselves in the context where we have the differences of only Poisson processes. It's pretty simple, one just have to condition the expectation of $E(|K|)$ for each case:

$$E(|K|) = \alpha(1-\delta) E(|K| \mid g) + \alpha\delta E(|K| \mid b) + (1-\alpha) E(|K| \mid n).$$

Then, remind $K = S - B$. S and B are, under the model assumption, Poisson processes describing the number of sells and buys in one day of trade. We only need two different kinds of Poisson processes to describe the mixture of Poisson processes resulting of informed and uninformed traders in each case ("good event", "bad event" and "no event"). Let's note them as follows $X_\epsilon^S \sim \mathcal{P}(\epsilon)$, $X_\epsilon^B \sim \mathcal{P}(\epsilon)$, $Y_\mu^B \sim \mathcal{P}(\mu)$ and $Y_\mu^S \sim \mathcal{P}(\mu)$, S and B labelling buys or sells. One finds then:

$$E(|K|) = \alpha(1-\delta) E\left(\left|X_\epsilon^S - \left(Y_\mu^B + X_\epsilon^B\right)\right|\right) + \alpha\delta E\left(\left|Y_\mu^B + X_\epsilon^B - X_\epsilon^S\right|\right)$$
$$+ (1-\alpha) E\left(\left|X_\epsilon^S - X_\epsilon^B\right|\right).$$

As all Poisson processes are independant one can sum them to produce new Poisson processes, as follows[3]:

$$E(|K|) = \alpha(1-\delta) E\left(\left|X_\epsilon - Y_{\mu+\epsilon}\right|\right) + \alpha\delta E\left(\left|Y_{\mu+\epsilon} - X_\epsilon\right|\right) + (1-\alpha) E\left(\left|X_\epsilon^{(1)} - X_\epsilon^{(2)}\right|\right).$$

One can thus sum the two first terms and obtain the following:

[3] S and B labels do not have any more importance, to differenciate Poisson processes of the last expectation we have thus just put label one and two to distinguish the "no event" case.

$$E\left(\left|K\right|\right)=\alpha E\left(\left|X_{\epsilon}-Y_{\mu+\epsilon}\right|\right)+\left(1-\alpha\right)E\left(\left|X_{\epsilon}^{(1)}-X_{\epsilon}^{(2)}\right|\right).$$

One has to treat finally two different cases:

- different intensities: first term
- same intensities: second term

3.2. How to Reach a First Order Approximate

In this subsection we will first see that main assumption to use Katti's result cannot be used to approximate the PIN. Therefore to approximate the PIN using authors' intuition we describe then the following two steps:

- one way to reach numerator exact value consists in using Ramasubban's ideas [18],
- first order asymptotic analysis involves separate cases to study sensitivity of the approximate to parameter's values.

3.2.1. Katti's Assumptions Are Not Met in the New Setting

We have seen that Katti's reference use the assumption that Poisson intensities are linked by a relation of the form $\lambda_1=\frac{\nu}{\lambda_2}$ where ν is independent of these parameters. Here the respective parameters would be $\mu+\epsilon$ and ϵ. The product $\epsilon\left(\epsilon+\mu\right)$ has clearly no single reason to be a constant. One could create some tricky cases, but it does not seem that the model would like to be limited to these cases (indeed, one may consider for example to fit the PIN parameters maximising likelihood, like in [1]). Thus the assumptions are not met and the reference [17] cannot be invoked to say $E\left|K\right|\approx\alpha\mu$ at first order, as it was done in [1] for example.

3.2.2. Computation of $E\left(\left|K\right|\right)$

Anyway, let's do nevertheless calculations to compute $E\left(\left|S-B\right|\right)$. We follow the same natural ideas of T. A. Ramasubban in this paper which treats only the case of same Poisson intensities [18]. We begin with:

$$\Delta_1=\alpha E\left(\left|X_{\epsilon}-Y_{\mu+\epsilon}\right|\right)+\left(1-\alpha\right)E\left(\left|X_{\epsilon}^{1}-X_{\epsilon}^{2}\right|\right).$$

Let's start with the easier calculation: the case where Poisson intensities are equal.

$$\begin{aligned}E\left(\left|X_{\epsilon}^{1}-X_{\epsilon}^{2}\right|\right)&=\sum_{i,j\in\mathbb{N}}P\left(X_{\epsilon}^{1}=i\right)P\left(X_{\epsilon}^{2}=j\right)\left|i-j\right|\\&=2\sum_{i\in\mathbb{N}}\sum_{j=0}^{i}P\left(X_{\epsilon}^{1}=i\right)P\left(X_{\epsilon}^{2}=j\right)\left(i-j\right)\\&=2\mathrm{e}^{-2\epsilon}\sum_{i\in\mathbb{N}}\sum_{j=0}^{i}\left(i-j\right)\frac{\epsilon^{i}}{i!}\frac{\epsilon^{j}}{j!}\\&=2\epsilon\mathrm{e}^{-2\epsilon}\sum_{i\in\mathbb{N}^*}\sum_{j=0}^{i}\frac{\epsilon^{i-1}}{\left(i-1\right)!}\frac{\epsilon^{j}}{j!}-\sum_{j=1}^{i}\frac{\epsilon^{i}}{i!}\frac{\epsilon^{j-1}}{\left(j-1\right)!}.\end{aligned}$$

All the sums separately exist, we can split them in two different ones:

$$
\begin{aligned}
E\left(\left|X_\epsilon^1 - X_\epsilon^2\right|\right) &= 2\epsilon \mathrm{e}^{-2\epsilon}\left(\sum_{i\in\mathbb{N}}\sum_{j=0}^{i+1}\frac{\epsilon^i}{i!}\frac{\epsilon^j}{j!} - \sum_{i\in\mathbb{N}^*}\sum_{j=0}^{i-1}\frac{\epsilon^i}{i!}\frac{\epsilon^j}{j!}\right)\\
&= 2\epsilon \mathrm{e}^{-2\epsilon}\left(\sum_{i\in\mathbb{N}}\frac{\epsilon^i}{i!}\left(\frac{\epsilon^{i+1}}{(i+1)!}+\frac{\epsilon^i}{i!}\right)\right)\\
&= 2\epsilon \mathrm{e}^{-2\epsilon}\left(\sum_{i\in\mathbb{N}}\frac{\epsilon^{2i+1}}{i!(i+1)!}+\sum_{i\in\mathbb{N}}\frac{\epsilon^{2i}}{i!^2}\right).
\end{aligned}
$$

One recognizes here a modified Bessel functions of first kind: for an integer n and, say scalar x, $I_n(x) = \sum_{i\in\mathbb{N}}\frac{\left(\frac{x}{2}\right)^{2i+n}}{i!(n+i)!}$. Here we obtain:

$$
E\left(\left|Y_B - Y_S\right|\right) = 2\epsilon \mathrm{e}^{-2\epsilon}\left(I_0(2\epsilon)+I_1(2\epsilon)\right).
$$

which is the result of Ramasubban's quoted paper. The computation with different intensities follow the same idea, expect that the symmetry of the two initial sums is broken, so we have to compute them separately.

$$
\begin{aligned}
E\left|X_\epsilon - Y_{\mu+\epsilon}\right| &= \sum_{i\in\mathbb{N}}\sum_{j=0}^{i} P\left(X_\epsilon = i\right)P\left(Y_{\epsilon+\mu} = j\right)(i-j)\\
&\quad + \sum_{i\in\mathbb{N}}\sum_{j=0}^{i} P\left(X_\epsilon = j\right)P\left(Y_{\epsilon+\mu} = i\right)(i-j).
\end{aligned}
$$

Let's calculate the first sum and then the second:

$$
\begin{aligned}
&\sum_{i\in\mathbb{N}}\sum_{j=0}^{i} P\left(X_\epsilon = i\right)P\left(Y_{\epsilon+\mu} = j\right)(i-j)\\
&= \mathrm{e}^{-2\epsilon-\mu}\left(\sum_{i=1}^{+\infty}\sum_{j=0}^{i}\frac{\epsilon^i}{(i-1)!}\frac{(\epsilon+\mu)^j}{j!} - \sum_{i=1}^{+\infty}\sum_{j=1}^{i}\frac{\epsilon^i}{i!}\frac{(\epsilon+\mu)^j}{(j-1)!}\right)\\
&= \mathrm{e}^{-2\epsilon-\mu}\left(\sum_{i=0}^{+\infty}\sum_{j=0}^{i+1}\frac{\epsilon^{i+1}}{i!}\frac{(\epsilon+\mu)^j}{j!} - \sum_{i=1}^{+\infty}\sum_{j=0}^{i-1}\frac{\epsilon^i}{i!}\frac{(\epsilon+\mu)^{j+1}}{j!}\right),
\end{aligned}
$$

which separates as follows as all sums exist separately:

$$
\begin{aligned}
&\sum_{i\in\mathbb{N}}\sum_{j=0}^{i} P\left(X_\epsilon = i\right)P\left(Y_{\epsilon+\mu} = j\right)(i-j)\\
&= \mathrm{e}^{-2\epsilon-\mu}\epsilon\sum_{i=0}^{+\infty}\frac{\epsilon^i}{i!}\left(\frac{(\epsilon+\mu)^{i+1}}{(i+1)!}+\frac{(\epsilon+\mu)^i}{i!}\right) - \mathrm{e}^{-2\epsilon-\mu}\mu\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\frac{\epsilon^{i+1}}{(i+1)!}\frac{(\epsilon+\mu)^j}{j!}.
\end{aligned}
$$

Replacing first sum of the rigth hand side by Bessel functions of second kind, we finally find:

$$
\begin{aligned}
&\sum_{i\in\mathbb{N}}\sum_{j=0}^{i} P\left(X_\epsilon = j\right)P\left(Y_{\epsilon+\mu} = i\right)(i-j)\\
&= \mathrm{e}^{-2\epsilon-\mu}\sqrt{\epsilon(\epsilon+\mu)}I_1\left(2\sqrt{\epsilon(\epsilon+\mu)}\right) + \mathrm{e}^{\epsilon(\epsilon+\mu)}\epsilon I_0\left(2\sqrt{\epsilon(\epsilon+\mu)}\right)\\
&\quad - \mathrm{e}^{-2\epsilon-\mu}\mu\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\frac{\epsilon^{i+1}}{(i+1)!}\frac{(\epsilon+\mu)^j}{j!}.
\end{aligned}
$$

For the second sum, we do an equivalent calculus and find the following:

$$\sum_{i\in\mathbb{N}}\sum_{j=0}^{i}P\left(X_{\epsilon}=j\right)P\left(Y_{\epsilon+\mu}=i\right)\left(i-j\right)$$
$$=\mathrm{e}^{-2\epsilon-\mu}\epsilon\sum_{i=0}^{+\infty}\frac{\left(\epsilon+\mu\right)^{i}}{i!}\left(\frac{\epsilon^{i+1}}{\left(i+1\right)!}+\frac{\epsilon^{i}}{i!}\right)-\mathrm{e}^{-2\epsilon-\mu}\mu\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\frac{\epsilon^{i+1}}{\left(i+1\right)!}\frac{\left(\epsilon+\mu\right)^{j}}{j!},$$

thus:

$$\sum_{i\in\mathbb{N}}\sum_{j=0}^{i}P\left(X_{\epsilon}=j\right)P\left(Y_{\epsilon+\mu}=i\right)\left(i-j\right)$$
$$=\mathrm{e}^{-2\epsilon-\mu}\frac{\epsilon^{2}}{\sqrt{\epsilon\left(\epsilon+\mu\right)}}I_{1}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)+\mathrm{e}^{-2\epsilon-\mu}\epsilon I_{0}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)$$
$$+\mathrm{e}^{-2\epsilon-\mu}\mu\sum_{i=0}^{+\infty}\frac{\left(\epsilon+\mu\right)^{i}}{i!}\sum_{j=0}^{i+1}\frac{\epsilon^{j}}{j!}.$$

If we put together all the terms we find:

$$E\left|X_{\epsilon}-Y_{\epsilon+\mu}\right|=\mathrm{e}^{-2\epsilon-\mu}\left(\frac{2\epsilon^{2}+\epsilon\mu}{\sqrt{\epsilon\left(\epsilon+\mu\right)}}I_{1}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)+2\epsilon I_{0}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)\right.$$
$$\left.-\mu\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\frac{\epsilon^{i+1}}{\left(i+1\right)!}\frac{\left(\epsilon+\mu\right)^{j}}{j!}+\mu\sum_{i=0}^{+\infty}\frac{\left(\epsilon+\mu\right)^{i}}{i!}\sum_{j=0}^{i+1}\frac{\epsilon^{j}}{j!}\right).$$

Arranging the last two sums of the left hand side of the equality we finnaly get:

$$E\left|X_{\epsilon}-Y_{\epsilon+\mu}\right|=\mathrm{e}^{-2\epsilon-\mu}\left(\frac{2\epsilon^{2}+\epsilon\mu}{\sqrt{\epsilon\left(\epsilon+\mu\right)}}I_{1}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)+2\epsilon I_{0}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)\right.$$
$$+\mu\sum_{i=0}^{+\infty}P\left(Y_{\epsilon+\mu}=i\right)\left(P\left(X_{\epsilon}\leq i+1\right)-P\left(X_{\epsilon}\geq i+1\right)\right).$$

Thus $E\left|K\right|$ equals:

$$E\left|K\right|=2\epsilon\left(1-\alpha\right)\mathrm{e}^{-2\epsilon}\left[I_{0}\left(2\epsilon\right)+I_{1}\left(2\epsilon\right)\right]$$
$$+\alpha\mathrm{e}^{-2\epsilon-\mu}\left[\frac{2\epsilon^{2}+\epsilon\mu}{\sqrt{\epsilon\left(\epsilon+\mu\right)}}I_{1}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)+2\epsilon I_{0}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}\right)\right]$$
$$+\alpha\mu\sum_{i=0}^{+\infty}P\left(Y_{\epsilon+\mu}=i\right)\left(P\left(X_{\epsilon}\leq i+1\right)-P\left(X_{\epsilon}\geq i+1\right)\right).$$

With an arbitrarily time length t for a trading period, we find:

$$E\left|K\right|=2\epsilon t\left(1-\alpha\right)\mathrm{e}^{-2\epsilon t}\left[I_{0}\left(2\epsilon t\right)+I_{1}\left(2\epsilon t\right)\right]$$
$$+\alpha\mathrm{e}^{-2\epsilon t-\mu t}\left[\frac{2\left(\epsilon t\right)^{2}+\epsilon\mu t^{2}}{\sqrt{\epsilon\left(\epsilon+\mu\right)}t}I_{1}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}t\right)+2\epsilon tI_{0}\left(2\sqrt{\epsilon\left(\epsilon+\mu\right)}t\right)\right]$$
$$+\alpha\mu t\sum_{i=0}^{+\infty}P\left(Y_{\left(\epsilon+\mu\right)t}=i\right)\left(P\left(X_{\epsilon t}\leq i+1\right)-P\left(X_{\epsilon t}\geq i+1\right)\right).$$

3.2.3. Analysis of the First Order Approximate

Recall that ϵ and μ are rates of uninformed and informed traders per day (in the original the PIN model). Thus, these parameters are pretty high integers: this is the first intuition behind first order approximate. Moreover, Hankel [19] derived an asymptotic expansion of modified Bessel function of first kind as follows:

$$I_\alpha(z) \sim \frac{\mathrm{e}^z}{\sqrt{2\pi z}}\left(1-\frac{4\alpha^2-1}{8z}+\frac{(4\alpha^2-1)(4\alpha^2-9)}{2!(8z)^2}\right.$$
$$\left.-\frac{(4\alpha^2-1)(4\alpha^2-9)(4\alpha^2-25)}{3!(8z)^3}+\cdots\right)$$

for $|z| \gg 1$ and $|\arg z| < \frac{\pi}{2}$

We first apply this expansion to $E|K|$ with the condition $\mu \gg 1$ and $\epsilon \gg 1$, as we consider there are a lot of informed and uninformed traders per day (compared to 1). We find the following:

$$E|K| \sim 2\sqrt{\frac{\epsilon}{\pi}}(1-\alpha)+\frac{\alpha}{2\sqrt{\pi}}\frac{\left(\epsilon+\sqrt{\epsilon(\epsilon+\mu)}\right)^2}{(\epsilon(\epsilon+\mu))^{\frac{3}{4}}}\mathrm{e}^{2\sqrt{\epsilon(\epsilon+\mu)}-2\epsilon-\mu}$$
$$+\alpha\mu\sum_{i=0}^{+\infty}P(Y_{\epsilon+\mu}=i)\left(P(X_\epsilon \le i+1)-P(X_\epsilon \ge i+1)\right).$$

Let's now distinguish these three cases:

- μ and ϵ are of same order,
- $\mu = o(\epsilon)$,
- $\epsilon = o(\mu)$,

If μ and ϵ are of same order,

in this case: $\sqrt{\epsilon(\epsilon+\mu)} < 2\epsilon-\mu$, thus one can neglect the corresponding term. We obtain:

$$E|K| \sim 2\sqrt{\frac{\epsilon}{\pi}}(1-\alpha)+\alpha\mu\sum_{i=0}^{+\infty}P(Y_{\epsilon+\mu}=i)\left(P(X_\epsilon \le i+1)-P(X_\epsilon \ge i+1)\right).$$

Thus $\text{PIN} \sim \frac{\alpha\mu}{\alpha\mu+2\epsilon}$ and

$$\frac{E|S-B|}{E(S+B)} \sim \frac{2\sqrt{\frac{\epsilon}{\pi}}(1-\alpha)+\alpha\mu\sum_{i=0}^{+\infty}P(Y_{\epsilon+\mu}=i)\left(P(X_\epsilon \le i+1)-P(X_\epsilon \ge i+1)\right)}{\alpha\mu+2\epsilon}.$$

if $\sqrt{\epsilon} \ll \mu$ then it reduces to:

$$E|K| \sim \alpha\mu\sum_{i=0}^{+\infty}P(Y_{\epsilon+\mu}=i)\left(P(X_\epsilon \le i+1)-P(X_\epsilon \ge i+1)\right).$$

Thus:

$$\frac{E|S-B|}{E(S+B)} \sim \frac{\alpha\mu}{\alpha\mu+2\epsilon}\sum_{i=0}^{+\infty}P(Y_{\epsilon+\mu}=i)\left(P(X_\epsilon \le i+1)-P(X_\epsilon \ge i+1)\right).$$

If $\mu = o(\epsilon)$,

we find:

$$\frac{E|S-B|}{E(S+B)} \sim \frac{2\sqrt{\frac{\epsilon}{\pi}}+\alpha\mu\sum_{i=0}^{+\infty}P(Y_{\epsilon+\mu}=i)\left(P(X_\epsilon \le i+1)-P(X_\epsilon \ge i+1)\right)}{2\epsilon}.$$

And: $\text{PIN} \sim \frac{\alpha\mu}{2\epsilon} \ll 1$.

If $\sqrt{\epsilon} \gg \mu$, then:

$$\frac{E|S-B|}{E(S+B)} \sim \frac{1}{\sqrt{\epsilon\pi}}.$$

If $\epsilon = o(\mu)$**,**

we find:

$$\frac{E|S-B|}{E(S+B)} \sim \sum_{i=0}^{+\infty} P(Y_{\epsilon+\mu} = i)\left(P(X_\epsilon \le i+1) - P(X_\epsilon \ge i+1)\right).$$

and $PIN \sim 1$

Thus, we can see that first order approximation depends a lot of:

- the respective values of μ and ϵ,
- and in a lot of cases of the weighted average of a given Poisson distribution of the difference between cumulative density functions from opposite parts of the tail of another Poisson distribution, *i.e.*: $\sum_{i=0}^{+\infty} P(Y_{\epsilon+\mu} = i)\left(P(X_\epsilon \le i+1) - P(X_\epsilon \ge i+1)\right)$

The first order approximation $E|S-B| \sim \alpha\mu$ proposed in [1] is not incorrect as we will see in the simulations, but sometimes, imprecise.

3.3. The Volume-Clock Paradigm: The Implicit Change of Model Assumptions

In this subsection, we describe the volume-clock framework and explain why its hypotheses lead to different results the PIN compared to the time-clock framework. More precisely, we first describe the new assumptions. Secondly, we make the computations within this new framework, which lead to a new value of the PIN.

3.3.1. The New Assumptions

In [3] D. Easley, M. de Prado and M. O'Hara describe a new model to compute easily the VPIN and therefore the PIN using the above previous results:

- $E(|S-B|) = \alpha\mu$, supposedly at first order,
- $E(S+B) = \alpha\mu + 2\epsilon$

They introduce the paradigm of volume clock and time bars. Let's first describe it and see that the assumptions are implicitly changed, but ignored. The idea is pretty simple. Consider a trade described by a time serie of price, say p_t, labelled with time t. First, They package trades in objects called "bars" that have a fixed time volume, *i.e.*: they aggregate the time serie in, for example, one-minute time bars. It is equivalent to a sampling of the time serie. Each bar is a kind of new trade with several rules to guess its price. Second they agreggate these time bars to form fixed in volum "buckets". Say these buckets have a volume V.

- Remark 1: nothing can ensure us that buckets will have a fixed volume size. Indeed, each time bar is sensitive to trading intensity. The last time bar can often be too big to be aggregated to a fixed size bucket. Which mean, that if one wants to force bucket size to be constant then, a lot of time bar won't be of one minute lenght. If one on the contrary wants to preserve time size to be constant, a lot of buckets might not be of constant volume size.

Suppose anyway that everything is ideal and that each bucket is of constant volume. Authors note τ the label of a bucket of volume V, V_τ, and V_τ^S and V_τ^B respectively the total number of sells and buys that occured in this bucket. They then refer to their previous work [1] result: $E\left(\left|V_\tau^S - V_\tau^B\right|\right) \approx \alpha\mu$. But even if the result does not hold as previously shown, one must note the following:

- First: here the bucket is constant in volume, thus filling volume time is random, it is a really strong hypothesis, as we have then: $V_\tau^S + V_\tau^B = V$ that holds almost surely,
- Second: they use the result indeed to say that as $V_\tau^S + V_\tau^B = V$ then the expectation equals $V = 2\epsilon + \alpha\mu$.
- But finally, one should remark that this equality lacks a time, as we are talking of rates of traders. In the first model, the time was one day, and implicitly one would multiply within the time-clock framework, rates by one day. Here, in the volume-clock framework, one does not control anymore time. One should take into account filling bucket time which is a new random variable. At first glance, the expression is inhomogenous and even if right, it is far from being trivial.

Indeed the authors preciss us "recall that we divide the trading day into equal-sized volume buckets and treat each volume bucket as equivalent to a period for information arrival". It's misleading. Recall that in the initial model time is fixed (one day) and thus volum is random. Here one has the contrary, volume is fixed and time is thus random. Let's detail a bit more the calculus with the new assumptions. To do so let's precise a bit more the new implicit framework.

3.3.2. A New Computation of $E\left(\left|V_t^S - V_t^B\right| \mid V\right)$

In fact we want to compute now $\Delta_1 = E\left(\left|V_\tau^S - V_\tau^B\right| \mid V\right)$, as bucket volume is fixed. Note $t - t'$ the filling time of the bucket τ and then note the following:

$$\Delta_1 = E\left(\left|S_t^{tot} - S_{t'}^{tot} - \left(B_t^{tot} - B_{t'}^{tot}\right)\right| \mid V\right),$$

with S_t^{tot}, $S_{t'}^{tot}$, B_t^{tot} and $B_{t'}^{tot}$ the Poisson processes of the total sell up to t and t' and the total buys up to t and t'. One has in distribution the following:

- $S_t^{tot} - S_{t'}^{tot} = N_{S,t-t'}^{tot}$
- $B_t^{tot} - B_{t'}^{tot} = N_{B,t-t'}^{tot}$

where $N_{S,t-t'}^{tot}$ and $N_{B,t-t'}^{tot}$ are Poisson processes describing total sells and buys in the bucket labelled with τ. The variables being independent, we can thus write the following:

$$\Delta_1 = E\left(\left|N_{S,t-t'}^{tot} - N_{B,t-t'}^{tot}\right| \mid V\right).$$

One must note that there is still the constraint of the volume of a bucket:

$$N_{S,t-t'}^{tot} + N_{B,t-t'}^{tot} = V.$$

Thus, imposing one value, imposes the other. Let's calculate Δ_1. First, one can condition the events: "good event" (*g*), "bad event" (*b*) and "no event" (*n*):

$$\Delta_1 = \alpha(1-\delta)E\left(\left|N^{tot}_{S,t-t'} - N^{tot}_{B,t-t'}\right| \mid g,V\right) + \alpha\delta E\left(\left|N^{tot}_{S,t-t'} - N^{tot}_{B,t-t'}\right| \mid b,V\right)$$
$$+(1-\alpha)E\left(\left|N^{tot}_{S,t-t'} - N^{tot}_{B,t-t'}\right| \mid n,V\right).$$

On each event, one knows the distribution of $N^{tot}_{S,t-t'}$ and $N^{tot}_{B,t-t'}$. One can then re-write it the following way[4]:

$$\Delta_1 = \alpha(1-\delta)E\left(\left|N^{tot,1}_{\epsilon(t-t')} - N^{tot,1}_{(\epsilon+\mu)(t-t')}\right| \mid V\right) + \alpha\delta E\left(\left|N^{tot,2}_{(\mu+\epsilon)(t-t')} - N^{tot,2}_{\epsilon(t-t')}\right| \mid V\right)$$
$$+(1-\alpha)E\left(\left|N^{tot,3}_{\epsilon(t-t')} - N^{tot,4}_{\epsilon(t-t')}\right| \mid V\right).$$

The two first terms corresponding to "good" or "bad events" are equal in distribution, that's why we have:

$$\Delta_1 = \alpha E\left(\left|N^{tot,1}_{\epsilon(t-t')} - N^{tot}_{(\epsilon+\mu)(t-t')}\right| \mid V\right) + (1-\alpha)E\left(\left|N^{tot,2}_{\epsilon(t-t')} - N^{tot,3}_{\epsilon(t-t')}\right| \mid V\right).$$

Before going further, let's implement the joint probability density function of for example, sells and buys and respective filling bucket time t-t' in the case of a bad event. Let's note it $f(S,B,t-t' \mid V,b)$. Now, we synthetise and refer to the great ideas of the proof of Kin and Le [14]. Remark first the following:

$$f(S,B,t-t' \mid V,b) = f(S,B \mid V,t-t',b)\, f(t-t' \mid V,b),$$

and as $f(V \mid t-t',b)$ follows a Poisson law of intensity $(t-t')(2\epsilon+\mu)$, then $f(t-t' \mid V,b)$ classically follows an Erlang law with the following parameters $\Gamma(t-t';V,2\epsilon+\mu)$. Second, as $S+B=V$ almost surely, we have the following equalities:

$$f(S,B \mid t-t',V,b) = f(S \mid t-t',V,b) = f(B \mid t-t',V,b),$$

and:

$$f(S,V \mid t-t',b) = f(B \mid V,t-t',b)\, f(V \mid t-t',b).$$

We know $f(V \mid t-t',b) \sim \mathcal{P}((2\epsilon+\mu)(t-t'))$ and considering for example a continuous bounded function *g*, one can guess easily $f(B \mid V,t-t',b)$ computing $E(g(S,V \mid t-t',b))$ using that $E(g(S,V \mid t-t',b)) = E(g(S,S+B \mid t-t',b))$. We find a binomial law $\mathcal{B}\left(S;V,\frac{\epsilon}{2\epsilon+\mu}\right)$ *i.e.:*

$$f(S \mid V,t-t',b) = f(B \mid V,t-t',b) = \mathcal{B}\left(S;V,\frac{\epsilon}{2\epsilon+\mu}\right) = \mathcal{B}\left(B;V,\frac{\epsilon+\mu}{2\epsilon+\mu}\right).$$

So finally:

[4] the label 1, 2, 3, … are used to note that these are the same distributions, but these are still different random variables.

$$f\left(S,B,t-t'\mid V,b\right)=\mathcal{B}\left(S;V,\frac{\epsilon}{2\epsilon+\mu}\right)\Gamma\left(V,2\epsilon+\mu\right).$$

The "no event" case is similar. We thus find the following:

$$f\left(S,B,t-t'\mid V\right)$$
$$=\alpha\mathcal{B}\left(S;V,\frac{1}{2}\right)\Gamma\left(t-t';V,2\epsilon\right)+\left(1-\alpha\right)\mathcal{B}\left(S;V,\frac{\epsilon}{2\epsilon+\mu}\right)\Gamma\left(t-t';V,2\epsilon+\mu\right),$$

And after an integration on the random variable t-t':

$$f\left(S,B\mid V\right)=\alpha\mathcal{B}\left(S;V\frac{1}{2}\right)+\left(1-\alpha\right)\mathcal{B}\left(S;V,\frac{\epsilon}{2\epsilon+\mu}\right).$$

Taking the previous joint probability into account we are thus computing the following expectations of let say X and Y in fact:

$$\Delta_1=\alpha E\left(\left|V-2X\right|\right)+\left(1-\alpha\right)E\left(\left|V-2Y\right|\right)$$

with $Y\sim\mathcal{B}\left(S;V,\frac{1}{2}\right)$ and $X\sim\mathcal{B}\left(V,S,\frac{\epsilon}{2\epsilon+\mu}\right)$

Moreover, if x follows the binomail distribution of which p.d.f is $\mathcal{B}\left(x;m,p\right)$, then using Jensen inequality for the concave function $y\to\sqrt{y}$ we have:

- $\frac{E\left(\left|m-2x\right|\right)}{m}=E\left(\left[\left(m-2x\right)^2\right]^{\frac{1}{2}}\right)\Big/m\le\sqrt{\left(2p-1\right)^2+\frac{4p\left(1-p\right)}{m}}\approx\left|2p-1\right|$ for large enough m and p differing from $\frac{1}{2}$
- $\frac{E\left(\left|m-2x\right|\right)}{m}\ge\frac{\left|E\left(m-2x\right)\right|}{m}=\left|2p-1\right|$.

Thus, for large enough V:

$$\Delta_1\approx V\alpha\left|2\frac{\epsilon}{2\epsilon+\mu}-1\right|+0,$$

i.e.

$$\Delta_1\approx V\frac{\alpha\mu}{2\epsilon+\mu}.$$

Thus the VPIN metric approximates the following for large enough n as shown by Kin and Le [14]:

$$\text{VPIN}=\frac{\sum_{\tau=1}^{n}\left|V_\tau^S-V_\tau^B\right|}{nV}\approx\frac{E\left(\left|V_\tau^S-V_\tau^B\right|\mid V\right)}{V}\approx\frac{\alpha\mu}{2\epsilon+\mu},$$

which is indeed different of $\text{PIN}=\frac{\alpha\mu}{\alpha\mu+2\epsilon}$.

3.4. Some Simulation Verification

We present here some simulation verification. First we present the framework and the experienced tested. Second, we present the results.

3.4.1. Framework and Experience Tested

For purpose of illustration, we compare the empirical form of $\frac{E(|B-S|)}{E(S+B)}$ with the PIN and the asymptotic limit[5] found within the time clock framework for different cases of ϵ and μ. It is pretty easy to do, as controlling ex-ante all the parameters of the model one then just has to generate the appropriate Poisson processes to obtain all the values. We illustrate the results with three examples:

- $\epsilon = o(\mu)$ and $\sqrt{\mu}$ of same order than ϵ: we took $\epsilon = 100$ and $\mu \in \{10000, 20000, 30000\}$,
- $\mu = o(\epsilon)$ and $\sqrt{\epsilon}$ of same order than μ: we took $\mu = 100$ and $\epsilon \in \{10000, 20000, 30000\}$,
- ϵ of same order than μ: we took[6] $\epsilon = 10000$ and $\mu \in \{10000, 2000, 30000\}$.

Remarks:

- We compute 20 values for each choice of ϵ and μ in the three cases above,
- For each of the 20 values, the empirical expectations are computed with an average of 10,000 values,
- To compute the sum $\sum_{i=0}^{+\infty} P(Y_{\epsilon+\mu} = i)(P(X_\epsilon \le i+1) - P(X_\epsilon \ge i+1))$, considering the values of ϵ and $\epsilon + \mu$, we have bounded the sum to $i = 100000$, when probability values starts to be then very little.

3.4.2. Results

On each case, we plot first the empirical numerator $E(|S-B|)$, $\alpha\mu$, and the asymptotic limit found (Figure 2, Figure 4 and Figure 6). Second, we plot $\frac{E(|B-S|)}{E(S+B)}$, the PIN (*i.e.* $\frac{\alpha\mu}{\alpha\mu + 2\epsilon}$ and the asymptotic limit divided by $\alpha\mu + 2\epsilon$ (Figure 3, Figure 5 and Figure 7).

Case1: $\epsilon = 100, \mu \in \{10000, 20000, 30000\}$

On Figure 2, first order and asymptotic estimations are very close.

Case 2: $\mu = 100, \epsilon \in \{10000, 20000, 30000\}$

On Figure 3 and Figure 4, one can see better the difference when one does not change μ anymore.

Case 3: $\epsilon = 10000, \mu \in \{10000, 2000, 30000\}$

This last case on Figure 6 and Figure 7 illustrates a market where the number of informed and uninformed traders are of same order.

[5] $$E|K| \sim 2\sqrt{\frac{\epsilon}{\pi}}(1-\alpha) + \frac{\alpha}{2\sqrt{\pi}} \frac{\left(\epsilon + \sqrt{\epsilon(\epsilon+\mu)}\right)^2}{\left(\epsilon(\epsilon+\mu)\right)^{\frac{3}{4}}} \mathrm{e}^{2\sqrt{\epsilon(\epsilon+\mu)} - 2\epsilon - \mu} .$$
$$+ \alpha\mu \sum_{i=0}^{+\infty} P(Y_{\epsilon+\mu} = i)(P(X_\epsilon \le i+1) - P(X_\epsilon \ge i+1)).$$

[6] This case is more tricky and actually the asymptotic limit is closer to the empirical value than the first order approximate proposed by the authors, but the trend is not obvious and need more study. We present here the good case that works fine. Further study must maybe be done.

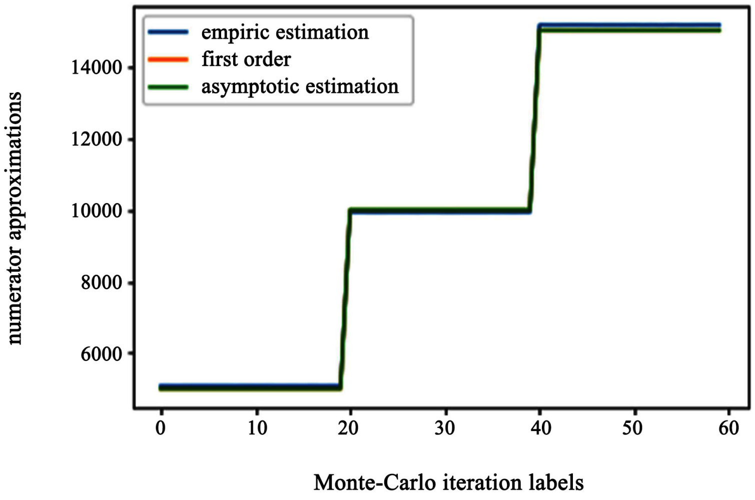

Figure 2. Empirical, asymptotic and first order numerators.

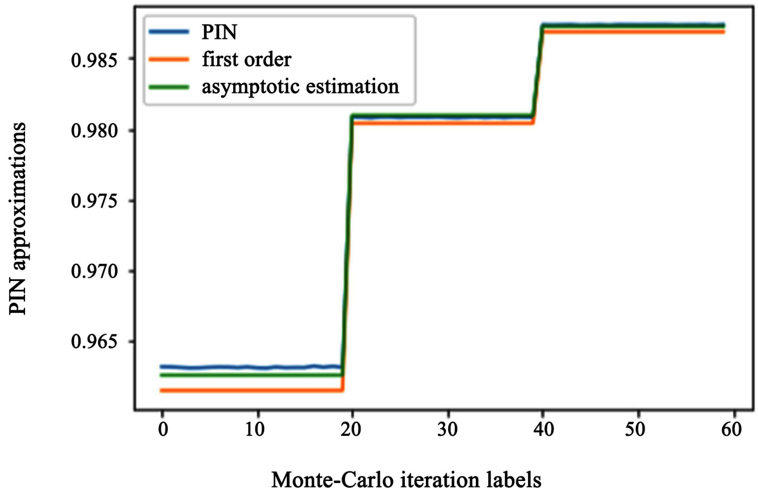

Figure 3. Empirical, asymptotic and first order approximations of the PIN.

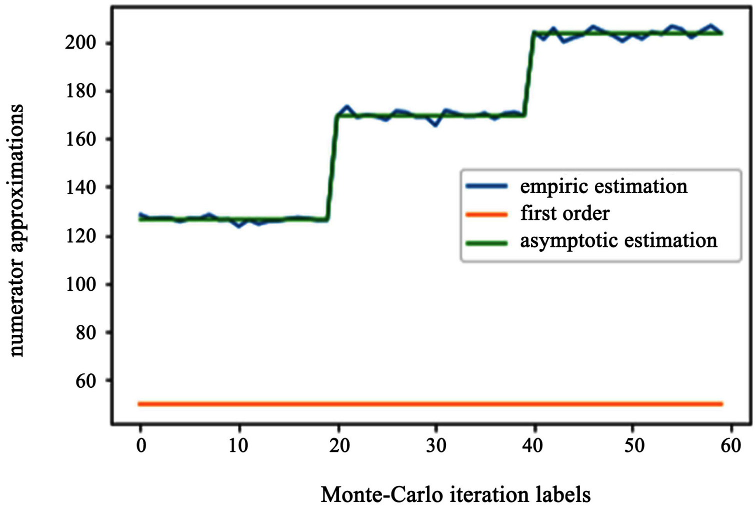

Figure 4. Empirical, asymptotic and first order numerators.

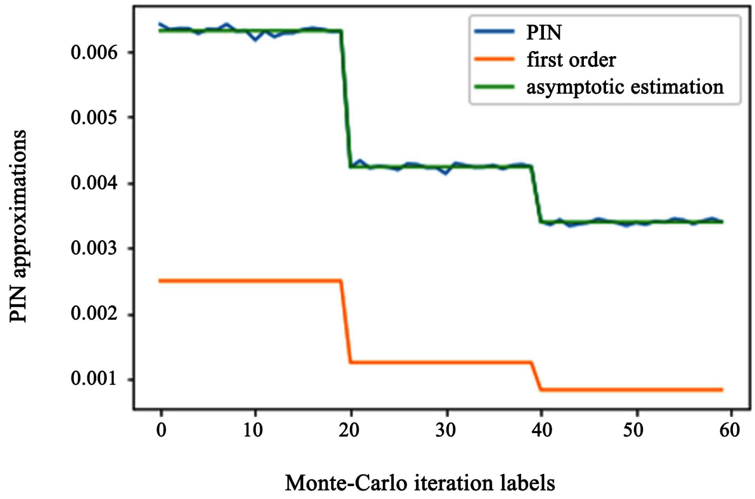

Figure 5. Empirical, asymptotic and first order approximations of the PIN.

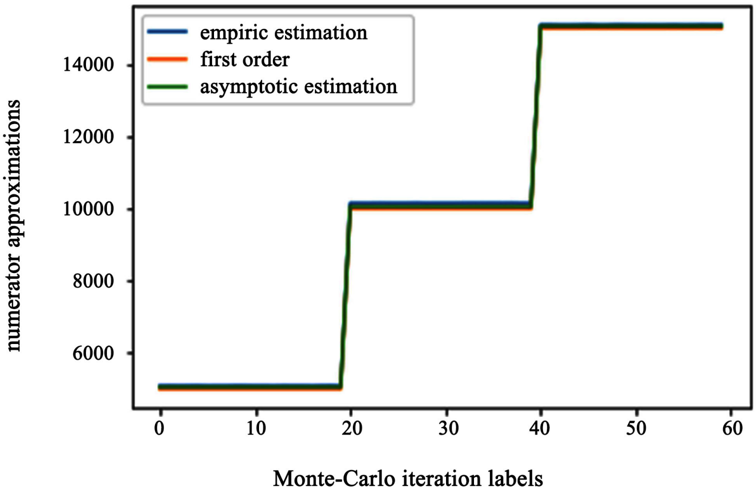

Figure 6. Empirical, asymptotic and first order numerators.

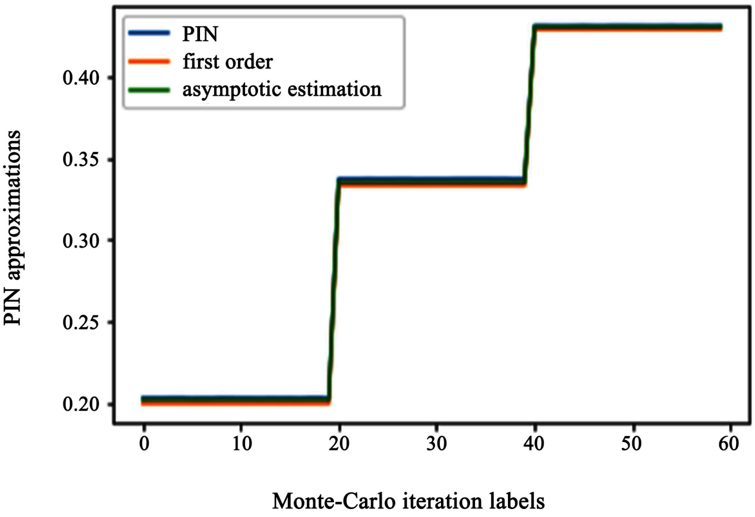

Figure 7. Empirical, asymptotic and first order approximations of the PIN.

4. Another Suggestion to Compute the PIN

In this section, we propose another way to compute the PIN. Indeed, as it was seen in the last section, the first order approximation of the PIN within the time-clock is not always precise and its theoretical foundation is not correct. Furthermore, the one we propose is only asymptotic and not easy to compute. Hence we propose an exact formula to compute the PIN in the time-clock framework. More precisely, in the first subsection we describe how to compute exactly the numerator $\alpha\mu$ and then the PIN. Secondly, we describe how numerically one can design at least one methodology to compute the PIN. Finally, we present some simulation verification of our results.

4.1. One PIN Upgrade

In this subsection, we detail how to compute exactly the PIN. Recall that the probability to obtain S sells and B buys during a period of length t is:

$$P\left(z_t=(S,B)\right)=(1-\delta)\alpha \mathrm{e}^{-(\mu+2\epsilon)t}\frac{\left((\mu+\epsilon)t\right)^B(\epsilon t)^S}{B!S!}+(1-\alpha)\mathrm{e}^{-2\epsilon t}\frac{(\epsilon t)^{B+S}}{B!S!}$$
$$+\alpha\delta \mathrm{e}^{-(\mu+2\epsilon)t}\frac{\left((\mu+\epsilon)t\right)^S\epsilon^B}{B!S!}.$$

Recall that to compute the PIN we have the assumption: $\delta=\frac{1}{2}$, thus we have:

$$P\left(z_t=(S,B)\right)=\frac{\alpha}{2}\mathrm{e}^{-(\mu+2\epsilon)t}\frac{\left((\mu+\epsilon)t\right)^B(\epsilon t)^S}{B!S!}+\frac{\alpha}{2}\mathrm{e}^{-(\mu+2\epsilon)t}\frac{\left((\mu+\epsilon)t\right)^S(\epsilon t)^B}{B!S!}$$
$$+(1-\alpha)\mathrm{e}^{-2\epsilon t}\frac{(\epsilon t)^{B+S}}{B!S!}.$$

So, if one notes $TT=S+B$ the total number of trades for this day, we find:

$$E(TT)=\frac{\alpha}{2}(\epsilon+\mu+\epsilon)t+\frac{\alpha}{2}(\mu+\epsilon+\epsilon)t+(1-\alpha)(\epsilon+\epsilon)t=(\alpha\mu+2\epsilon)t,$$

and we even have:

$$E(S)=E(B)=\left(\epsilon+\frac{\alpha\mu}{2}\right)t=\frac{E(TT)}{2}.$$

So to estimate the PIN denominator, one can first use for an arbitrary time period an average of S, B or TT. Let's work with S and take a time period of length t. Let's estimate the numerator $\frac{\alpha\mu t}{2}$. To do this, we firstly explicit the margin probability function to obtain S sells in a time period of length t and secondly we compute its first three moments. Thirdly we explain how to compute α and hence the numerator, which finally leads to a new PIN formula.

4.1.1. Margin Function

The probability to obtain S sells during a time period of length t is the following:

$$\begin{aligned}P\left(z_t=S\right)&=\frac{\alpha}{2}\mathrm{e}^{-\epsilon t}\frac{\left(\epsilon t\right)^S}{S!}+\left(1-\alpha\right)\mathrm{e}^{-\epsilon t}\frac{\left(\epsilon t\right)^S}{S!}+\frac{\alpha}{2}\mathrm{e}^{-\left(\mu+\epsilon\right)t}\frac{\left(\left(\epsilon+\mu\right)t\right)^S}{S!}\\&=\left(1-\frac{\alpha}{2}\right)\mathrm{e}^{-\epsilon t}\frac{\left(\epsilon t\right)^S}{S!}+\frac{\alpha}{2}\mathrm{e}^{-\left(\mu+\epsilon\right)t}\frac{\left(\left(\epsilon+\mu\right)t\right)^S}{S!}.\end{aligned}$$

4.1.2. Computation of First Three Moments

Let's compute the moment-generating function of this process. We will estimate the numerator using relations between moments. Let u be a real value, let V_S be the random variable representing the volume of sells and t the fixed time period associated. We have:

$$E\left(\mathrm{e}^{V_S u}\right)=\left(1-\frac{\alpha}{2}\right)\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}+\frac{\alpha}{2}\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)}.$$

Let's compute the first three moments of V_S :

- First moment:

$$E\left(V_S\mathrm{e}^{V_S u}\right)=\left(1-\frac{\alpha}{2}\right)\epsilon t\mathrm{e}^u\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}+\frac{\alpha}{2}\left(\epsilon+\mu\right)t\mathrm{e}^u\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)},$$

so:

$$E\left(V_S\right)=\left(\epsilon+\frac{\alpha\mu}{2}\right)t.$$

- Second moment:

$$\begin{aligned}E\left(V_S^2\mathrm{e}^{V_S u}\right)=&\left(1-\frac{\alpha}{2}\right)\epsilon t\mathrm{e}^u\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}+\left(1-\frac{\alpha}{2}\right)\left(\epsilon t\right)^2\mathrm{e}^{2u}\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}\\&+\frac{\alpha}{2}\left(\epsilon+\mu\right)t\mathrm{e}^u\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)}+\frac{\alpha}{2}\left(\left(\epsilon+\mu\right)t\right)^2\mathrm{e}^{2u}\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)},\end{aligned}$$

so:

$$E\left(V_S^2\right)=\left(\epsilon+\frac{\alpha\mu}{2}\right)t+\left(\alpha\mu\epsilon+\epsilon^2+\frac{\alpha\mu^2}{2}\right)t^2,$$

i.e. we have the classic decomposition:

$$E\left(V_S^2\right)=E\left(V_S\right)+E\left(V_S\left(V_S-1\right)\right).$$

- Third moment:

$$\begin{aligned}E\left(V_S^3\mathrm{e}^{V_S u}\right)=&\left(1-\frac{\alpha}{2}\right)\epsilon t\mathrm{e}^u\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}+\left(1-\frac{\alpha}{2}\right)\left(\epsilon t\right)^2\mathrm{e}^{2u}\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}\\&+2\left(1-\frac{\alpha}{2}\right)\left(\epsilon t\right)^2\mathrm{e}^{2u}\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}+\left(1-\frac{\alpha}{2}\right)\left(\epsilon t\right)^3\mathrm{e}^{3u}\mathrm{e}^{\epsilon t\left(\mathrm{e}^u-1\right)}\\&+\frac{\alpha}{2}\left(\epsilon+\mu\right)t\mathrm{e}^u\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)}+\frac{\alpha}{2}\left(\left(\epsilon+\mu\right)t\right)^2\mathrm{e}^{2u}\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)}\\&+2\frac{\alpha}{2}\left(\left(\epsilon+\mu\right)t\right)^2\mathrm{e}^{2u}\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)}+\frac{\alpha}{2}\left(\left(\epsilon+\mu\right)t\right)^3\mathrm{e}^{3u}\mathrm{e}^{\left(\epsilon+\mu\right)t\left(\mathrm{e}^u-1\right)},\end{aligned}$$

so:

$$E\left(V_S^3\right)=\left(\epsilon+\frac{\alpha\mu}{2}\right)t+3\left(\alpha\mu\epsilon+\epsilon^2+\frac{\alpha\mu^2}{2}\right)t^2+\left(\epsilon^3+\frac{3}{2}\alpha\mu\epsilon^2+\frac{3}{2}\alpha\epsilon\mu^2+\frac{\mu^3}{2}\alpha\right)t^3,$$

i.e. we just wrote:

$$E\left(V_S^3\right)=E\left(V_S\right)+3E\left(V_S\left(V_S-1\right)\right)+E\left(V_S\left(V_S-1\right)\left(V_S-2\right)\right).$$

4.1.3. Estimation of α

Remark the following:

$$\begin{aligned}
&E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2\\
&=\left(\alpha\mu\epsilon+\epsilon^2+\frac{\alpha\mu^2}{2}-\left(\epsilon+\frac{\alpha\mu}{2}\right)^2\right)t^2\\
&=\left(\frac{\mu^2\alpha}{2}-\frac{\alpha^2\mu^2}{4}\right)t^2=\left(\frac{\alpha\mu t}{2}\right)^2\left[\frac{2-\alpha}{\alpha}\right].
\end{aligned}$$

Then with the same idea let's compute the following:

$$\begin{aligned}
&E\left(V_S\left(V_S-1\right)\left(V_S-2\right)\right)-\left(E\left(V_S\right)\right)^3\\
&=\left[\left(\epsilon^3+\frac{3}{2}\alpha\mu\epsilon^2+\frac{3}{2}\alpha\epsilon\mu^2+\frac{\mu^3}{2}\alpha\right)-\left(\epsilon+\frac{\alpha\mu}{2}\right)^3\right]t^3\\
&=3\epsilon t\left(\frac{\alpha\mu t}{2}\right)^2\left[\frac{2-\alpha}{\alpha}\right]+\left(\frac{\alpha\mu t}{2}\right)^3\left[\frac{4-\alpha^2}{\alpha^2}\right]\\
&=3\epsilon t\left(\frac{\alpha\mu t}{2}\right)^2\left[\frac{2-\alpha}{\alpha}\right]+\left(\frac{\alpha\mu t}{2}\right)^2\frac{2-\alpha}{\alpha}\frac{\alpha\mu t}{2}\frac{\alpha+2}{\alpha},
\end{aligned}$$

and we know that $\epsilon t=E\left(V_S\right)-\frac{\alpha\mu t}{2}$ and that $\left(\frac{\alpha\mu t}{2}\right)^2\frac{2-\alpha}{\alpha}=E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2$, so:

$$\begin{aligned}
&E\left(V_S\left(V_S-1\right)\left(V_S-2\right)\right)-\left(E\left(V_S\right)\right)^3\\
&=\left(E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2\right)\left(3E\left(V_S\right)-\frac{3\alpha\mu t}{2}+\frac{\mu t}{2}\left(2+\alpha\right)\right),
\end{aligned}$$

i.e.

$$\frac{E\left(V_S\left(V_S-1\right)\left(V_S-2\right)\right)-\left(E\left(V_S\right)\right)^3}{E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2}=3E\left(V_S\right)+\frac{\alpha\mu t}{2}\frac{2\left(1-\alpha\right)}{\alpha},$$

If we use again the formula, we can then replace $\frac{\alpha\mu t}{2}$ by $\sqrt{\frac{E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2}{\frac{2-\alpha}{\alpha}}}$:

$$\begin{aligned}
&\frac{E\left(V_S\left(V_S-1\right)\left(V_S-2\right)\right)-\left(E\left(V_S\right)\right)^3}{\left(E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2\right)}\\
&=3E\left(V_S\right)+\sqrt{\frac{E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2}{\frac{2-\alpha}{\alpha}}}\frac{2\left(1-\alpha\right)}{\alpha},
\end{aligned}$$

so:

$$\frac{\dfrac{E\left(V_S\left(V_S-1\right)\left(V_S-2\right)\right)-\left(E\left(V_S\right)\right)^3}{E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2}-3E\left(V_S\right)}{\sqrt{E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2}}=\frac{2\left(1-\alpha\right)}{\sqrt{\alpha\left(2-\alpha\right)}}.$$

If we arrange a bit the expression on denominator and numerator on the left hand side of the equation, we remark the following:

- Remark 1:

$$E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2=Var\left(V_S\right)-E\left(V_S\right),$$

- Remark 2:

$$E\left(V_S\left(V_S-1\right)\left(V_S-2\right)\right)-\left(E\left(V_S\right)\right)^3-3E\left(V_S\right)E\left(V_S^2\right)+3E\left(V_S\right)^2+3\left(E\left(V_S\right)\right)^3$$
$$=E\left(\left(V_S-E\left(V_S\right)\right)^3\right)-3Var\left(V_S\right)+2E\left(V_S\right).$$

Thus:

$$\frac{E\left(\left(V_S-E\left(V_S\right)\right)^3\right)-3Var\left(V_S\right)+2E\left(V_S\right)}{\left(Var\left(V_S\right)-E\left(V_S\right)\right)^{\frac{3}{2}}}=\frac{2\left(1-\alpha\right)}{\sqrt{\alpha\left(2-\alpha\right)}}.$$

Introducing the skewness γ and the following notations: $\sigma=\sqrt{Var\left(V_S\right)}$, $\gamma=E\left(\left(\frac{V_S-E\left(V_S\right)}{\sigma}\right)^3\right)$ and $m=E\left(V_S\right)$, we obtain finally:

$$\frac{\gamma\sigma^3-3\sigma^2+2m}{\left(\sigma^2-m\right)^{\frac{3}{2}}}=\frac{2\left(1-\alpha\right)}{\sqrt{\alpha\left(2-\alpha\right)}}.$$

Skewness, standard deviation and expectation are measured from data. To estimate α we thus just have to solve the following second order equation on α:

$$\alpha^2-2\alpha+\frac{4}{4+\dfrac{\left(\gamma\sigma^3-\sigma^2+2m\right)^2}{\left(\sigma^2-m\right)^3}}=0.$$

The discriminant is positive: $\Delta=4\left(1-\dfrac{4}{4+\dfrac{\left(\gamma\sigma^3-\sigma^2+2m\right)^2}{\left(\sigma^2-m\right)^3}}\right)$. As α is a probability, we finally find:

$$\alpha=1-\sqrt{1-\frac{4}{4+\dfrac{\left(\gamma\sigma^3-\sigma^2+2m\right)^2}{\left(\sigma^2-m\right)^3}}},$$

which is indeed between 0 and 1.

4.1.4. Estimation of $\frac{\alpha\mu t}{2}$

We know that $E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2=\left(\frac{\alpha\mu t}{2}\right)^2\left[\frac{2-\alpha}{\alpha}\right]$, so let's replace the α of the right hand side of the equality (not with μt) by previous expresion. We then estimate $\frac{\alpha\mu t}{2}$. We finally obtain the following as $E\left(V_S\left(V_S-1\right)\right)-\left(E\left(V_S\right)\right)^2=\sigma^2-m$ with the previous notations:

$$\frac{\alpha\mu t}{2}=\sqrt{\frac{1-\sqrt{1-\frac{4}{4+\frac{\left(\gamma\sigma^3-\sigma^2+2m\right)^2}{\left(\sigma^2-m\right)^3}}}}{1+\sqrt{1-\frac{4}{4+\frac{\left(\gamma\sigma^3-\sigma^2+2m\right)^2}{\left(\sigma^2-m\right)^3}}}}}\left(\sigma^2-m\right)$$

4.1.5. A New PIN Formula

Finally we obtain the following equivalent exact formula:

$$\text{PIN}=\frac{\frac{\alpha\mu t}{2}}{\epsilon t+\frac{\alpha\mu t}{2}},$$

i.e.:

$$\text{PIN}=\frac{1}{m}\sqrt{\frac{1-\sqrt{1-\frac{4}{4+\frac{\left(\gamma\sigma^3-\sigma^2+2m\right)^2}{\left(\sigma^2-m\right)^3}}}}{1+\sqrt{1-\frac{4}{4+\frac{\left(\gamma\sigma^3-\sigma^2+2m\right)^2}{\left(\sigma^2-m\right)^3}}}}}\left(\sigma^2-m\right),$$

or after simplifying a bit:

$$\text{PIN}=\frac{2}{m}\frac{\left(\sigma^2-m\right)^2}{\sqrt{4\left(\sigma^2-m\right)^3+\left(\gamma\sigma^3-3\sigma^2+2m\right)^2}+\gamma\sigma^3-3\sigma^2+2m}.$$

One then just have to estimate on a arbitrary time lenght *t*, *m*, σ and γ to estimate the PIN number. The difficulty is then put on estimating on this time period the volume of direction of trades. We describe further a possible framework to compute this number. One can verify numerically that these two formula give the exact same numbers of the PIN.

4.2. A New Framework to Compute the PIN

In this subsection we explain, how at least one framework can be designed to compute the PIN. We would like to compute the PIN number from time, let say t, and period length, let say η, *i.e.* from t to $t+\eta$. With previous framework, we obviously have:

$$\text{PIN}_{t,t+\eta} = \frac{\frac{\alpha_{t,t+\eta}\mu_{t,t+\eta}t}{2}}{\epsilon_{t,t+\eta}t + \frac{\alpha_{t,t+\eta}\mu_{t,t+\eta}t}{2}},$$

as, all numbers α, μ and ϵ are defined on these time t and period η. And we also have:

$$\text{PIN}_{t,t+\eta} = \frac{2}{m}\frac{\left(\sigma^2 - m\right)^2}{\sqrt{4\left(\sigma^2 - m\right)^3 + \left(\gamma\sigma^3 - 3\sigma^2 + 2m\right)^2} + \gamma\sigma^3 - 3\sigma^2 + 2m},$$

where, m, γ and σ are calculated for the volume of sell $V_{S,t,t+\eta}$, between t and $t+\eta$.

Thus two things must be implemented to well estimate the PIN:

- the empirical averages implicitly behind m, σ and γ: we will have to put some hypothesis on the time series of volumes to use classic theorems.
- the volume of sells: one needs a model of classifier to guess on a given amount of time the number of sells within the total volume of sells.

Estimation of m, σ and γ

We would like to use the law of large number. We basically need random variable independant and identically distributed. Here: noting $N_{S,t}$ the Poisson process of sells at time t (*i.e.* the number of sells at t). Then we have:

$$V_{S,t,t+\eta} = N_{S,t+\eta} - N_{S,t}.$$

According to the model, Nature chooses at each time period η the parameters and independently each day. So $\left(V_{S,t,t+\eta}\right)_t$ is a sequence of (successive non-overlapping) in dependant random variables. But, the $V_{S,t,t+\eta}$ are not identically distributed. Nothing guarantees it. Indeed, Nature's choices won't necessarily be the same, and so $\alpha_{t,t+\eta}$, $\gamma_{t,t+\eta}$ and $\sigma_{t,t+\eta}$. To handle with this, one can do the following. We need a statistically significant mean. Within the time period: $[t,t+\eta[$ [7], Nature's choice is the same, so considering n intervals of length $\frac{\eta}{n}$ within $[t,t+\eta[$, the random variables $\left(V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}}\right)_{i=1,\cdots,n}$ are then independent and identically distributed. For n high enough the following approximations hold:

- $m \approx \frac{1}{n}\sum_{i=1}^{n} V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}}$,
- $\sigma \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}} - \frac{1}{n-1}\sum_{i=1}^{n} V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}}\right)^2}$,

[7] Let's suppose that choices are made in this time interval, to not bother about possibly overlapping Nature's choice.

- $\gamma \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}}-\frac{1}{n}\sum_{i=1}^{n}V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}}-\frac{1}{n-1}\sum_{i=1}^{n}V_{S,t+\frac{(i-1)\eta}{n},t+\frac{i\eta}{n}}\right)^{2}}}\right)^{3}$,

Thus the choices to make here are:

- the time length η,
- the number n of sub-intervals to have a precise average.

To reduce standard variation of $\text{PIN}_{t,t+\eta}$, one direct way to do it is to take both the averages of the PIN estimated using volume of sells (let's note it now the $\text{PIN}_{t,t+\eta}^{S}$ and the PIN estimated using volume of buys (let's note it $\text{PIN}_{t,t+\eta}^{V}$). Indeed, the previous calculations are exactly the same if one would have use volume of buys instead of sells. And within the PIN framework V_S and V_B are independent random variables. So:

$$\text{PIN}_{t,t+\eta} = \frac{1}{2}\left(\text{PIN}_{t,t+\eta}^{S} + \text{PIN}_{t,t+\eta}^{V}\right),$$

and so, an estimate of $\text{PIN}_{t,t+\eta}^{S}$ is only a function of the process (V_S) and $\text{PIN}_{t,t+\eta}^{V}$ the same function but depending of the process (V_B). Thus, these two estimates being independent, if one notes σ_{PIN_S} the standard deviation using only the process V_S, now the standard deviation σ_{PIN} with both process equals the following:

$$\sigma_{\text{PIN}} = \frac{\sigma_{\text{PIN}_S}}{\sqrt{2}}.$$

4.3. Some Simulation Verification

We present finally some simulation verification. First we describe its framework. Second we present the results. The values of parameter tested are exactly the same as in the last framework, as we would like to compare previous results with the values of our new formula. The only difference which slightly change our framework, is that to compute the new formula one needs more sample. We detail it now.

4.3.1. Framework and Experience Tested

For purpose of illustration, we compare the empirical form $\frac{E\left(|B-S|\right)}{E\left(S+B\right)}$ with the PIN and the new formula[8] found within the time clock framework for dif-

[8] We use in these simulations for symmetry reasons this formula.

$$\frac{1}{m}\sqrt{\frac{1-\sqrt{1-\frac{4}{4+\frac{\left(\gamma\sigma^{2}-\sigma^{2}+2m\right)^{2}}{\left(\sigma^{2}-m\right)^{3}}}}}{1+\sqrt{1-\frac{4}{4+\frac{\left(\gamma\sigma^{2}-\sigma^{2}+2m\right)^{2}}{\left(\sigma^{2}-m\right)^{3}}}}}}$$

ferent cases of ϵ and μ. It is pretty easy to do, as controlling ex-ante all the parameters of the model one then just has to generate the appropriate Poisson processes to obtain all the values. We illustrate the results with three examples:

- $\epsilon = o(\mu)$ and $\sqrt{\mu}$ of same order than ϵ: we took $\epsilon = 100$ and $\mu \in \{10000, 20000, 30000\}$,
- $\mu = o(\epsilon)$ and $\sqrt{\epsilon}$ of same order than μ: we took $\mu = 100$ and $\epsilon \in \{10000, 20000, 30000\}$,
- ϵ of same order than μ: we took[9] $\epsilon = 10000$ and $\mu \in \{10000, 2000, 30000\}$.

Remarks:

- We compute 20 values for each choice of ϵ and μ in the three cases above,
- For each of the 20 values, for a choice of ϵ and μ, we generate 1,000,000 Poisson processes, we divide them in 100 consecutive intervals of 10,000 values. For each of the 100 intervals we compute empirical average to approximate mean m, standard deviation σ and skewness γ. We then compute an approximation of the PIN with an average of these 100 values[10].

4.3.2. Results

On each case (Figure 8, Figure 9 and Figure 10), we plot $\frac{E(|B-S|)}{E(S+B)}$ (VPIN), the PIN ($\frac{\alpha\mu}{\alpha\mu + 2\epsilon}$) and the new PIN value (labelled as NPIN).

Case1: $\epsilon = 100, \mu \in \{10000, 20000, 30000\}$

On Figure 8, new formula (NPIN) and PIN are very close.

Case 2: $\mu = 100, \epsilon \in \{10000, 20000, 30000\}$

Here on Figure 9 one can see better the difference when one does not change μ anymore.

Case 3: $\epsilon = 10000, \mu \in \{10000, 2000, 30000\}$

This last case on Figure 10 illustrates a market where the number of informed and uninformed traders are of same order. The VPIN really slightly over-estimates the true PIN value.

In any case one sees that new formula estimated is closer than the VPIN one. By the way, we have checked that new PIN formula obviously equals true PIN formula for any parameter ϵ, μ and α of the model.

5. Conclusions

In this last section, we present first a general summary of our findings. Then we propose suggestion for further research on this topic.

[9] This case is more tricky and actually asumptotic limit is closer to the empirical value than first order approximate proposed by authors, but the tren is not obvious and needs more study. We present here the good case that works fine. Further study must maybe be done.

[10] This double average equals traditional the VPIN formula as values are consecutive.

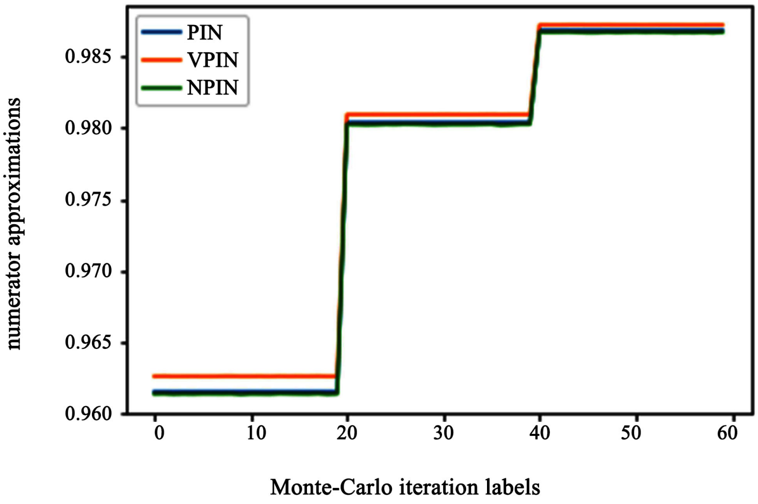

Figure 8. Old (VPIN) and new approximation (NPIN) of the PIN.

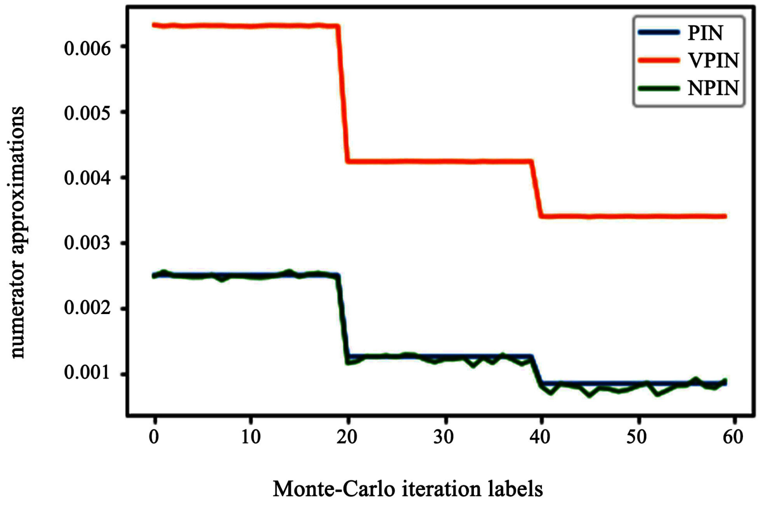

Figure 9. Old (VPIN) and new approximation (NPIN) of the PIN.

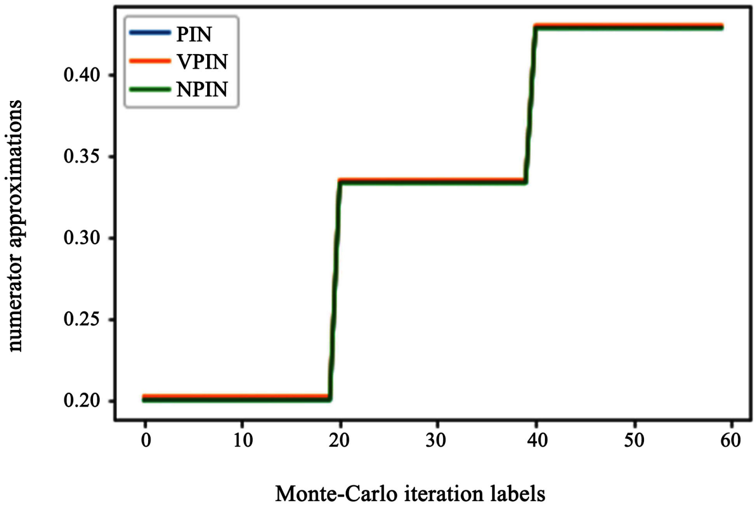

Figure 10. Old (VPIN) and new approximation (NPIN) of the PIN.

In this study we have analyzed the theoretical foundation of the PIN model and we have shown that its time-clock framework makes it hard to apply the VPIN original heuristic to estimate the probability of informed trading. Indeed, first order asymptotic is not that simple to estimate theoretically and in practice. That's why we propose another way to estimate the PIN, which is theoretically exact and hence more precise than the asymptotic formula, which is confirmed by our first tests. Moreover, the study recalls and highlights the difference of the volume-clock and time-clock paradigms which leads to a different formula of the PIN, and which respective hypotheses cannot therefore be used simultaneously to approximate the PIN.

Here are some ideas to further study this precise subject:

- test and compare the performance of the new formula within the time-clock framework with real trading data: find local optima parameters (n, η, trade classification algorithm, …) to maximize prediction quality,
- analyze and assess stability of the new formula and compare it to other ones.

Acknowledgements

We thank the Editor and the referee for their comments. Useful guidance and discussions in the LBNL team are gratefully acknowledged.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Easley, D., Engle, R.F., O'Hara, M. and Wu, L. (2008) Time-Varying Arrival Rates of Informed and Uninformed Trades. *Journal of Financial Econometrics*, **6**, 171-207. https://doi.org/10.1093/jjfinec/nbn003

[2] Easley, D., de Prado, M.L. and O'Hara, M. (2012) The Volume Clock: Insights into the High Frequency Paradigm. *Journal of Portfolio Management*, **39**, 19-29.

[3] Easley, D., de Prado, M.L. and O'Hara, M. (2012) Flow Toxicity and Liquidity in a High Frequency World. *Review of Financial Studies*, **25**, 1457-1493. https://doi.org/10.1093/rfs/hhs053

[4] Easley, D., de Prado, M.L. and O'Hara, M. (2011) The Microstructure of the "Flash Crash": Flow Toxicity, Liquidity, Crashes and the Probability of Informed Trading. *The Journal of Portfolio Management*, **37**, 118-128. https://doi.org/10.2139/ssrn.1695041

[5] Zheng, Y.M. (2017) VPIN and the China's Circuit-Breaker. *International Journal of Economics and Finance*, **9**, 126. https://doi.org/10.5539/ijef.v9n12p126

[6] Abad, D. (2011) From PIN to VPIN: An Introduction to Order Flow Toxicity. *The Spanish Review of Financial Economics*, **6**, 8-13. https://doi.org/10.1016/j.srfe.2012.10.002

[7] Wu, K.S., *et al.* (2013) A Big Data Approach to Analyzing Market Volatility. *Algorithmic Finance*, **2**, 241-267.

[8] Easley, D., de Prado, M.L. and O'Hara, M. (2011) The Exchange of Flow Toxicity.

The Journal of Trading, **6**, 8-13. https://doi.org/10.3905/jot.2011.6.2.008

[9] Andersen, T.G. and Bondarenko, O. (2014) VPIN and the Flash Crash. *The Journal of Financial Markets*, **17**, 1-46. https://doi.org/10.1016/j.finmar.2013.05.005

[10] Andersen, T.G. and Bondarenko, O. (2014) Reflecting on the VPIN Dispute. *Journal of Financial Markets*, **17**, 53-64. https://doi.org/10.1016/j.finmar.2013.08.002

[11] Andersen, T.G. and Bondarenko, O. (2015) Assessing Measures of Order Flow Toxicity and Early Warning Signals for Market Turbulence. *Review of Finance*, **19**, 1-54. https://doi.org/10.1093/rof/rfu041

[12] Abad, D., Massot, M. and Pascual, R. (2017) Evaluating VPIN as a Trigger for Single-Stock Circuit Breakers. *Journal of Banking and Finance*, **86**, 21-36. https://doi.org/10.1016/j.jbankfin.2017.08.009

[13] Pöppe, T., Moss, S. and Schiereck, D. (2016) The Sensitivity of VPIN to the Choice of Trade Classification Algorithm. *Journal of Banking and Finance*, **73**, 165-181. https://doi.org/10.1016/j.jbankfin.2016.08.006

[14] Ke, W.-C. and Lin, H.-W.W. (2017) An Improved Version of the Volume-Synchronized Probability of Informed Trading. *Critical Finance Review*, **6**, 357-376.

[15] Easley, D., de Prado, M.L. and O'Hara, M. (2017) An Improved Version of the Volume-Synchronized Probability of Informed Trading (VPIN): A Comment. *Critical Finance Review*, **6**, 377-379.

[16] Easley, D., Kiefer, N.M., O'Hara, M. and Paperman, J.B. (1996) Liquidity, Information, and Infrequently Traded Stocks. *Journal of Finance*, **51**, 1405-1436. https://doi.org/10.2307/2329399

[17] Katti, S.K. (1959) Moments of the Absolute Difference and Absolute Deviation of Discrete Distributions. *The Annals of Mathematical Statistics*, **31**, 78-85. https://doi.org/10.1214/aoms/1177705989

[18] Ramasubban, T.A. (1959) The Mean Difference and the Mean Deviation of Some Discontinuous Distributions. *Biometrika*, **45**, 549-556. https://doi.org/10.1093/biomet/45.3-4.549

[19] Abramowitz, M., Stegun, I.A., Abramowitz, M. and Stegun, I.A. (1964) Handbook of Mathematical Functions. National Bureau of Standards Applied Mathematics Series 55, 377-378.