INSIDER TRADING IN BRAZIL'S STOCK MARKET

A PREPRINT

Thiago Marzagão

June 29, 2021

ABSTRACT

How much insider trading happens in Brazil's stock market? Previous research has used the model proposed by Easley et al. [1996] to estimate the probability of insider trading (PIN) for different stocks in Brazil. Those estimates have a number of problems: i) they are based on a factorization that biases the PIN downward, especially for high-activity stocks; ii) they fail to account for boundary solutions, which biases most PIN estimates upward (and a few of them downward); and iii) they are a decade old and therefore based on a very different market (for instance, the number of retail investors grew from 600 thousand in 2011 to 3.5 million in 2021). In this paper I address those three problems and estimate the probability of insider trading for 431 different stocks in the Brazilian stock market, for each quarter from October 2019 to March 2021.

Keywords: insider trading, asset pricing, financial econometrics
JEL codes: G14, G12, C58

Introduction

How much insider trading happens in Brazil's stock market?¹² Easley et al. [1996] offer a model we can use to estimate the probability of insider trading (PIN) for different stocks. That model only requires buy and sell orders. The PIN has been estimated for many different stocks in many different markets, including the Brazilian one - see the work of Barbedo et al. [2010], Martins and Paulo [2013], and Martins et al. [2013]. As the next sections will show, current estimates of the PIN for Brazilian stocks are biased. Moreover, they are a decade old - the most recent ones refer to 2011. In this paper I address these problems and estimate the PIN for each of 431 different stocks in the Brazilian stock market, for each quarter from October 2019 to March 2021.

Model

In the model proposed by Easley et al. [1996] and subsequently updated by Easley et al. [2010] - henceforth EHO -, an information event for a given asset occurs on day t with probability α . The information event is negative (bad news) with probability δ and positive (good news) with probability $1 - \delta$. Buy orders from uninformed traders arrive at a rate ε_b and sell orders from uninformed traders arrive at a rate ε_s . If an information event happens the arrival rate of orders from informed traders (sell orders if bad news, buy orders if good news) is μ . Any information, good or bad, is known only to the informed traders; the market maker and the other traders never observe it. Figure 1, adapted from Easley et al. [1996], summarizes the model.

Each arrival rate - μ , ε_b , and ε_s - follows a separate Poisson process, independent from the others. Also, these Poisson processes are independent for each day. The market maker is a Bayesian agent who knows the probabilities associated with each branch; each day he uses the observed arrival of buy and sell orders to update those priors. These updates determine the bid and ask prices the market maker sets.

¹ Author's contact and ORCID information: tmarzagao@gmail.com (<https://orcid.org/0000-0003-0395-3985>). All views expressed here are the author's.

² I thank Fernando Sola, Fernando Vassoler, Nelson Oliveira, Eduardo Paiva, David Cosac, and André Rocha for helpful comments. All errors are mine.

Figure 1: The trading process. Adapted from Easley et al. [1996].

The joint probability distribution of the parameter vector $\theta = \{\alpha, \mu, \delta, \varepsilon_b, \varepsilon_s\}$, with B_t and S_t being the number of buy orders and sell orders on day t , respectively, is given by³

$$\begin{aligned}
 f(B_t, S_t | \theta) = & \alpha \delta e^{(-\varepsilon_b)} \frac{\varepsilon_b^{B_t}}{B_t!} e^{-(\varepsilon_s + \mu)} \frac{(\varepsilon_s + \mu)^{S_t}}{S_t!} \\
 & + \alpha(1 - \delta) e^{-(\varepsilon_b + \mu)} \frac{(\varepsilon_b + \mu)^{B_t}}{B_t!} e^{(-\varepsilon_s)} \frac{\varepsilon_s^{S_t}}{S_t!} \\
 & + (1 - \alpha) e^{(-\varepsilon_b)} \frac{\varepsilon_b^{B_t}}{B_t!} e^{(-\varepsilon_s)} \frac{\varepsilon_s^{S_t}}{S_t!}
 \end{aligned} \tag{1}$$

Intuitively, this probability distribution is a mixture of the Poisson processes for the three arrival rates, weighted by the probability of a day with bad news ($\alpha\delta$), of a day with good news ($\alpha(1 - \delta)$), and of a day without an information event ($1 - \alpha$). (See EHO, p. 296.)

We can obtain the estimated parameter vector $\hat{\theta} = \{\hat{\alpha}, \hat{\mu}, \hat{\delta}, \hat{\varepsilon}_b, \hat{\varepsilon}_s\}$ by maximizing the likelihood of (1) with the constraints that $\alpha, \delta \in [0, 1]$ and $\mu, \varepsilon_b, \varepsilon_s \in [0, \infty)$. Because the days are independent from each other, the likelihood of observing $(B_t, S_t)_{t=1}^T$ over T days is simply the product of the daily likelihoods:

$$\mathcal{L}((B_t, S_t)_{t=1}^T | \theta) = \prod_{t=1}^T \mathcal{L}(\theta | B_t, S_t) \tag{2}$$

³In what follows I draw heavily from EHO and from Tiniç and Celik [2017].

We can then estimate the PIN for a given asset in a given period as

$$\widehat{PIN} = \frac{\hat{\alpha}\hat{\mu}}{\hat{\alpha}\hat{\mu} + \hat{\varepsilon}_b + \hat{\varepsilon}_s} \quad (3)$$

Factorization-related bias

Finding the $\hat{\theta}$ vector that maximizes the likelihood of $(B_t, S_t)_{t=1}^T$ is computationally difficult for high-activity stocks as that results in equation (1) having very large denominators and exponents. For this reason EHO propose the following rearrangement of the likelihood function:

$$\begin{aligned} \mathcal{L}_{EHO}(\theta|B_t, S_t) = & \ln [\alpha \delta e^{-\mu} x_b^{B_t-M_t} x_s^{-M_t} + \alpha(1-\delta) e^{-\mu} x_b^{-M_t} x_s^{S_t-M_t} + (1-\alpha) x_b^{B_t-M_t} x_s^{S_t-M_t}] \\ & + B_t \ln(\varepsilon_b + \mu) + S_t \ln(\varepsilon_s + \mu) - (\varepsilon_b + \varepsilon_s) + M_t [\ln(x_b) + \ln(x_s)] - \ln(S_t! B_t!) \end{aligned} \quad (4)$$

where $M_t = \min(B_t, S_t) + \max(B_t, S_t)/2$, $x_b = \varepsilon_b/(\mu + \varepsilon_b)$, and $x_s = \varepsilon_s/(\mu + \varepsilon_s)$.

While EHO is computationally easier than the likelihood originally proposed by Easley et al. [1996], Lin and Ke [2011] show that it underestimates the PIN for high-activity stocks. They review several EHO applications and find a downward bias in 44% of the PIN estimates. To address this bias Lin and Ke [2011] propose an alternative factorization (henceforth LK):

$$\begin{aligned} \mathcal{L}_{LK}(\theta|B_t, S_t) = & \ln [\alpha \delta e^{(e_{1t}-e_{maxt})} + \alpha(1-\delta) e^{(e_{2t}-e_{maxt})} + (1-\alpha) e^{(e_{3t}-e_{maxt})}] \\ & + B_t \ln(\varepsilon_b + \mu) + S_t \ln(\varepsilon_s + \mu) - (\varepsilon_b + \varepsilon_s) + e_{maxt} - \ln(S_t! B_t!) \end{aligned} \quad (5)$$

where $e_{1t} = -\mu - B_t \ln(1 + \mu/\varepsilon_b)$, $e_{2t} = -\mu - S_t \ln(1 + \mu/\varepsilon_s)$, $e_{3t} = -B_t \ln(1 + \mu/\varepsilon_b) - S_t \ln(1 + \mu/\varepsilon_s)$, and $e_{maxt} = \max(e_{1t}, e_{2t}, e_{3t})$.

As I discuss in detail in the methods section, I estimated the PIN for 431 stocks in the Brazilian stock market from October 2019 to March 2021, using both the EHO and LK factorizations. Comparing the two sets of results shows that EHO estimates are indeed lower than LK estimates (average of 16.2% vs average of 22.2% respectively) and that they are especially lower for high-activity stocks. The difference between the (natural log of) LK estimates and EHO estimates correlates 0.6 ($p < 0.01$) with trading activity. Figure 2 shows how the LK-EHO difference grows as trading activity grows.

None of the peer-reviewed papers that apply Easley et al. [1996]'s model to the Brazilian stock market - Barbedo et al. [2010], Martins and Paulo [2013], and Martins et al. [2013] - addresses the downward bias identified by Lin and Ke [2011]. Except for Barbedo et al. [2010], which predates both EHO and LK, they all use the EHO factorization.

Boundary-related bias

Maximizing $\mathcal{L}(\theta|B_t, S_t)$, be it with the LK factorization or with the EHO factorization, must obey the constraints that $\alpha, \delta \in [0, 1]$ and $\mu, \varepsilon_b, \varepsilon_s \in [0, \infty)$. Yan and Zhang [2006] show that when the maximization solution falls on the boundary of those constraints - for example, when $\hat{\alpha} = 0$ or $\hat{\alpha} = 1$ -, the estimates can be severely biased. In one of their examples they show that with boundary solutions a PIN of 0.131 can be erroneously estimated as low as 0.000 or as high as 0.801. In general, when $\hat{\alpha} = 0$ there will be a downward bias and when $\hat{\alpha} = 1$ there will be an upward bias. When $\hat{\alpha} = 0$ the PIN is necessarily zero, so in the end the non-zero PIN estimates will have an upward bias.

To overcome this bias Yan and Zhang [2006] propose a grid search algorithm for finding good initial values for the estimation of (4). To find these initial values the algorithm tries 125 different combinations of parameters and chooses the combination that maximizes (4), after discarding all the combinations that produced negative arrival rates (negative μ , ε_b , or ε_s) and all the combinations that failed to converge.

While Yan and Zhang [2006]'s algorithm (henceforth YZ) does seem to improve the estimates it is computationally intensive - it requires that the PIN for each stock be estimated 125 times. As Gan et al. [2015] note, "It is not uncommon for market microstructure academics to spend weeks or even months estimating PIN parameters." (p. 3).

Gan et al. [2015] propose an alternative for finding initial values that does not rely on grid search and is therefore much faster. They propose an alternative algorithm (henceforth GAN) for finding initial values. The first step of GAN is to compute $X_t = B_t - S_t$, the difference between buy orders and sell orders on day t . We then use a hierarchical agglomerative clustering (HAC) algorithm to sequentially cluster X_t , bottom-up, based on the distance

Figure 2: LK-EHO differences vs trading activity. Each dot represents a stock-quarter ($n = 2245$).

function $D(I, J) = |X_i - X_j| \forall i \neq j$, until three clusters are formed. The cluster with the highest $\bar{X}_t$ is labeled as "high signal" (H), the cluster with the lowest $\bar{X}_t$ is labeled "low signal" (L) and the remaining cluster is labeled "no signal" (N). Let w_c be the proportion of samples in cluster c , such that $\sum_{c=1}^3 w_c = 1$. The initial values to be used in the maximum likelihood estimation are then $\hat{\alpha} = w_L + w_H$, $\hat{\delta} = w_L / \hat{\alpha}$, $\hat{\varepsilon}_b = \frac{w_L}{w_L + w_N} \bar{B}_L + \frac{w_L}{w_L + w_N} \bar{B}_N$, $\hat{\varepsilon}_s = \frac{w_H}{w_H + w_N} \bar{S}_H + \frac{w_N}{w_H + w_N} \bar{B}_N$, and $\hat{\mu} = \frac{w_H}{w_L + w_H} (\bar{B}_H - \hat{\varepsilon}_b) + \frac{w_L}{w_L + w_H} (\bar{S}_H - \hat{\varepsilon}_s)$.

None of the peer-reviewed papers that apply Easley et al. [1996]'s model to the Brazilian stock market addresses the bias resulting from boundary solutions, be it through YZ, GAN, or any other method.

Data and method

I collected transaction-level data (i.e., trade tick data) for each stock that was traded in Brazil's stock market (B3) at least once in the period between 2019-10-01 and 2021-03-31. I only included stocks, both ordinary and preferred; therefore I left out ETFs, Brazilian Depositary Receipts (BDRs), and unit trusts. I collected the data from the platform MetaTrader, using its Python API.⁴

The data contains, for each tick, its timestamp (down to the millisecond) and its aggressor side, i.e., whether the transaction was a buy (a buyer sends a buy market order that closes someone's sell limit order) or a sell (a seller sends a sell market order that closes someone's buy limit order).⁵ This departs from previous research, most of which did not have aggressor side information and had to rely on Lee and Ready [1991]'s algorithm to infer whether a tick was a buy or a sell.

⁴See <http://thiagomarzagao.com/2021/05/12/mt5/> for details on how to scrape data from MetaTrader.

⁵A small number of ticks - around 1% of the total - had both the buy and sell flags, which means a buy market order and a sell market order were sent simultaneously to the same broker. I discarded these cases.

I aggregated the ticks for each trading day and then estimated the PIN for each stock-quarter comprehended between 2019-10-01 and 2021-03-31.⁶ As mentioned before, I produced two sets of estimates: one using EHO and one using LK. In both cases, to find the initial values I used the GAN algorithm as implemented in the R package *InfoTrad* developed by Tiniç and Celik [2017]. The EHO estimates were meant only to check whether they would be lower than the LK estimates, which I already discussed; hence from now on I will only refer to the LK estimates, which are more accurate.

In total there were 2416 stock-quarters with trading activity. For 171 of them the trading activity was too low to allow estimation and the result was an R exception.⁷ Hence in the end a total of 2245 PIN estimates were produced. These 2245 stock-quarters refer to 431 different stocks.

Results

The average PIN for the 2245 stock-quarters is 22.2%. Figure 3 shows its distribution. All the estimates and code are publicly available.⁸

Figure 3: Distribution of PIN estimates. Each dot represents a stock-quarter ($n = 2245$).

Table 1 shows some descriptive statistics for PIN the estimates, globally and per quarter. It also shows how the PIN correlates with trading activity (number of ticks) and it breaks down the PIN estimates per level of governance: Novo

⁶The choice of quarterly PIN estimates is mainly so that the results are comparable to those of previous research, as discussed in further detail in a later section.

⁷Also, not every stock was traded in every quarter. To give an example, PETZ3 only IPO'ed in September 2020.

⁸<https://github.com/thiagomarzagao/pin/>

Mercado (NM), which is the strictest one (companies are not allowed to issue preferred stocks, for instance, and have to follow a number of transparency and compliance requirements)⁹, versus all other levels^{10, 11}.

Table 1: Descriptive statistics of PIN estimates.

	mean	median	25th percentile	75th percentile	correlation w/ trading activity	NM	non-NM	n
2019Q4	22.3%	19.9%	13.1%	29.6%	-0.42	16.9%	25.6%	340
2020Q1	22.8%	19.4%	14.7%	28.2%	-0.33	16.5%	26.6%	347
2020Q2	21.8%	18.3%	12.6%	28.1%	-0.40	14.5%	26.5%	350
2020Q3	21.1%	18.4%	12.7%	27.7%	-0.42	14.3%	24.8%	388
2020Q4	22.5%	19.5%	15.1%	27.3%	-0.36	17.3%	25.5%	403
2021Q1	22.5%	19.5%	15.2%	28.2%	-0.40	17.6%	25.5%	417
all quarters	22.2%	19.4%	13.9%	28.2%	-0.37	16.2%	25.7%	2245

It is reassuring that these statistics are stable across all quarters, as wild fluctuations would suggest the possibility of computational problems. Also, the negative (and in all cases statistically significant with $p < 0.001$) correlation of PIN with trading activity corroborates Easley et al. [1996]’s empirical finding that high-activity stocks have lower PIN. Finally, it is also reassuring that the average PIN is lower for companies that operate in Novo Mercado, as they are required to follow a number of strict compliance and transparency rules.¹²

Another sign that our PIN estimates are probably not too far off the mark is that they correlate with certain regulatory indicators. The Brazilian counterpart to the US Securities and Exchange Commission is the *Comissão de Valores Mobiliários* - CVM. CVM continually monitors the market and when it detects anomalous activity with some stock price - for instance, unusual volatility or a spike in trading volume - it asks the issuing company about potential causes. I downloaded all such inquiries for the period under analysis (2010-10-01 to 2021-03-31).¹³ The total number of these inquiries correlates 0.36 with the average PIN of each company ($p < 0.001$).¹⁴ While these inquiries are not specifically about insider trading (the anomalous activity may be due to other reasons, like market manipulation), it is reassuring that they have a positive correlation with PIN.

Comparison with previous research

The average PIN of 22.2% found here is in line with previous findings for Brazil’s stock market. Martins and Paulo [2013], with a sample of 229 stocks from 2010-2011, find an average PIN of 24.9%. Martins et al. [2013], with a sample of 198 stocks from 2011, find an average PIN of 22.9%. The big difference is with Barbedo et al. [2010], who find an average PIN of 12.5%. But Barbedo et al. [2010]’s sample has only 48 stocks and they do not use any factorization (which introduces computational problems, as discussed before); hence their estimates are probably not comparable to the other ones.¹⁵

⁹The full list of requirements is available at B3’s website: http://www.b3.com.br/pt_br/produtos-e-servicos/solucoes-para-emissores/segmentos-de-listagem/sobre-segmentos-de-listagem/

¹⁰The PIN does not vary substantially among these other governance levels, so for brevity I do not discuss them here and instead just aggregate them under the label "non-Novo Mercado".

¹¹When the company entered or (more rarely) left the Novo Mercado in the middle of the quarter I assigned it the governance level it had in the previous quarter. In a small number of cases no governance level information was available; in these cases I inputted non-NM.

¹²I also looked into how the PIN varies across economic sectors, using the sector each company declares as its primary sector for fiscal purposes. But the number of stocks per sector is too low to yield any meaningful comparisons. I used the National Registry of Economic Activities (CNAE) at the three-digit level and only three sectors have at least 100 stock-quarters in the sample. On average each sector has only 21.3 stock-quarters in the sample. At lower levels of sector aggregation (four- or five-digit levels) that average is even lower.

¹³The inquiries are available at http://dados.cvm.gov.br/dataset/cia_aberta-doc-ipe

¹⁴The average PIN across all stock-quarters, for each company.

¹⁵Agudelo et al. [2015] also estimate PIN values for Brazilian stocks, but their estimates cannot be compared to the other ones discussed here as the underlying model is fundamentally different. Here, as well as in Martins and Paulo [2013] and in Martins et al. [2013], I treat the arrival rates of informed and uninformed traders as constant, in line with Easley et al. [1996] and with EHO, whereas Agudelo et al. [2015] treat them as time-varying. That enables them to estimate the PIN for stock-days - as opposed to stock-quarters, which is what most PIN-related research does -, but we cannot compare daily and quarterly probabilities. Moreover, unlike here and unlike Martins and Paulo [2013] and Martins et al. [2013], Agudelo et al. [2015] do not allow the arrival rate to differ between uninformed buyers and uninformed sellers.

Both Martins and Paulo [2013] and Martins et al. [2013] use the EHO factorization and neither corrects for boundary-related bias. As discussed before, EHO biases the PIN downwards, especially for high-activity stocks, and boundary issues bias non-zero PIN estimates upwards. I used LK factorization, which does not have the downward bias of EHO, and I used GAN to produce initial values, which corrects for boundary issues. Yet we all obtained similar PIN averages. Hence it is possible that both sources of bias are cancelling each other out in previous research. Checking for that possibility would require inspecting the PIN distribution in each of those previous works and comparing those distributions to the one produced here. That is not possible though, as both of those papers provide only aggregate statistics.

Another possibility is that the similar PIN averages are due to increased trading activity since 2011. Martins and Paulo [2013] and Martins et al. [2013] both use data from a decade ago, when Brazil's stock market had about 600 thousand retail investors. Now it has almost 3.5 million retail investors. Easley et al. [1996]'s central empirical finding is that the PIN is lower for frequently traded stocks. Hence it is possible that applying EHO and without correcting for boundary problems we would find a lower average PIN for the 2019-2021 period studied here. That however would not provide conclusive evidence in one direction or another: the EHO downward bias is worse for high-activity stocks, so a lower PIN using EHO and 2019-2021 might simply be the result of a larger bias.

Yet another possibility is that in previous PIN estimates the samples were biased. Here I used every stock that was traded at least once (i.e., that had at least a single tick) at any moment between 2019-10-01 and 2021-03-31. Martins and Paulo [2013] and Martins et al. [2013], on the other hand, only included stocks that were traded at least once *in each trading day* in the periods they studied, which means about 60 days per quarter. They justify that choice based on Easley et al. [1996]'s assertion that "[P]revious research has shown that ... a sixty day trading window is sufficient to allow reasonably precise estimation of the parameters" (p. 1416). That 60-day guideline, however, predates the development of improved estimation procedures, like the LK factorization and the GAN algorithm, both of which produce more precise estimates, and both of which I use here. And in any case, dropping stock-quarters with fewer than 60 trading days does not change the average PIN dramatically here - it goes from 22.2% to 18.6%.

It is also possible that, in previous research, the upward bias introduced into non-zero PIN estimates by boundary problems dominates the downward EHO bias. If that is the case than most of the estimates in Martins and Paulo [2013] and in Martins et al. [2013] are biased upward and therefore the true average PIN has increased since 2011. Again, however, we cannot know whether that is the case.

Alternatively, it is possible that the GAN-based estimates produced here are themselves biased. Ersan and Alici [2016] - henceforth EA - argue that GAN's clustering procedure introduces a downward bias into the PIN estimates. EA propose an alternative clustering procedure that they claim fixes the problem. I applied EA's procedure to all stock-quarters in my sample, to see how that would change the results, but the average EA-based PIN turned out to be lower than the average GAN-based PIN: 20% vs 22.2% respectively. I used the same data and factorization (LK) in both cases, so if EA were right we would expect the average EA-based PIN to be higher, not lower, than the average GAN-based PIN.

Moreover, the GAN algorithm was able to generate PIN estimates for 2245 of the 2416 of the stock-quarters in the sample whereas the EA algorithm generated PIN estimates for only 1626 of them. That happened because EA initially clusters each $|X_t|$ into only two clusters, "news" and "no news", and only later splits the "news" cluster into a "good news" one and a "bad news" one. In many cases the initial "news" cluster had fewer than two samples, which meant it could not be partitioned into "good news" and "bad news"; the result was an R exception. One would expect this to happen with low-activity stocks but there was no relation with stock activity. To give one example, the problem happened even with the stock-quarter ABEV3-2020Q1, which has over two million trades.

Interestingly, if a given stock-quarter has only 0-2 samples in the "news" cluster we would imagine its PIN to be zero or close to zero, as there can be no informed traders if there is no information to be traded upon. If that initial "news"/"no news" clustering step is correct then EA is fundamentally flawed: it prevents estimation for low-PIN cases. More likely though that first step is producing nonsensical clusters, as the alternative explanation would be an implausibly high number of zero-PIN stock-quarters - 790, or 32% of the total, which is too high for the concentrated, low-volume Brazilian market.

In sum, it looks like EA's findings, which are all based on simulated data, do not necessarily apply to real-world data. Clustering the samples into "good news"/"bad news"/"no news" clusters right away, as GAN does, seems to yield better results.

Finally, when it comes to governance levels the results here depart markedly from previous research. Martins and Paulo [2013] and Martins et al. [2013] find almost identical PIN averages across all levels of governance. They look into each specific governance level below Novo Mercado, which prevents a direct comparison - here I only compare Novo Mercado vs non-Novo Mercado -, but the largest difference they find is the one between Novo Mercado and Nível 1, in Martins et al. [2013]: 22.3% vs 24.2% respectively. Here, in contrast, the average PIN is 16.2% for Novo Mercado

and 25.7% for non-Novo Mercado - a difference of almost ten percentage points, and one that is consistent across all quarters. Hence unlike in previous research, here I find that in Novo Mercado investors are better protected against insider trading.¹⁶

Future directions

Recent years have seen a number of developments in PIN research. For instance, Albuquerque et al. [2020] develop a Bayesian implementation that allows the researcher to incorporate expert opinions. It might be interesting to compare the estimates produced by the conventional (frequentist) model with the estimates produced by a Bayesian approach with informative priors. Another direction would be to estimate the PIN for other asset classes like options and cryptocurrencies. It would also be interesting to do a more detailed analysis of PIN estimates for the so called "meme stocks" (like \$GME and \$AMC) that have recently become the object of pump-and-dump schemes organized on Reddit forums and other social media. Finally, it would be interesting to know whether the model is applicable to bets on prediction markets like PredictIt.

References

- Diego A. Agudelo, Santiago Giraldo, and Edwin Villarraga. Does pin measure information? informed trading effects on returns and liquidity in six emerging markets. *International Review of Economics and Finance*, 39:149–161, 2015. ISSN 1059-0560. doi: <https://doi.org/10.1016/j.iref.2015.04.002>. URL <https://www.sciencedirect.com/science/article/pii/S1059056015000659>.
- Pedro Albuquerque, Cibele da Silva, Leonardo Bosque, Eduardo Nakano, and Yaohao Peng. Probability of informed trading: a bayesian approach. *International Journal of Applied Decision Sciences*, 13:1, 01 2020. doi: 10.1504/IJADS.2020.10023337.
- Claudio Barbedo, Eduardo Camilo-da Silva, and Ricardo Leal. Premium listing segments and information based trading in brazil. *Academia*, pages 1–19, 01 2010.
- David Easley, Nicholas M. Kiefer, Maureen O’Hara, and Joseph B. Paperman. Liquidity, information, and infrequently traded stocks. *The Journal of Finance*, 51(4):1405–1436, 1996. doi: <https://doi.org/10.1111/j.1540-6261.1996.tb04074.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb04074.x>.
- David Easley, Soeren Hvidkjaer, and Maureen O’Hara. Factoring information into returns. *The Journal of Financial and Quantitative Analysis*, 45(2):293–309, 2010. ISSN 00221090, 17566916. URL <http://www.jstor.org/stable/27801486>.
- Oguz Ersan and Aslı Alici. An unbiased computation methodology for estimating the probability of informed trading (pin). *Journal of International Financial Markets, Institutions and Money*, 43:74–94, 2016. ISSN 1042-4431. doi: <https://doi.org/10.1016/j.intfin.2016.04.001>. URL <https://www.sciencedirect.com/science/article/pii/S104244311630021X>.
- Quan Gan, Wang Chun Wei, and David Johnstone. A faster estimation method for the probability of informed trading using hierarchical agglomerative clustering. *Quantitative Finance*, 15(11):1805–1821, 2015. doi: 10.1080/14697688.2015.1023336. URL <https://doi.org/10.1080/14697688.2015.1023336>.
- Charles M. C. Lee and Mark J. Ready. Inferring trade direction from intraday data. *The Journal of Finance*, 46(2): 733–746, 1991. doi: <https://doi.org/10.1111/j.1540-6261.1991.tb02683.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1991.tb02683.x>.
- Hsiou-Wei Lin and Wen-Chyan Ke. A computing bias in estimating the probability of informed trading. *Journal of Financial Markets*, 14(4):625–640, 2011. ISSN 1386-4181. doi: <https://doi.org/10.1016/j.finmar.2011.03.001>. URL <https://www.sciencedirect.com/science/article/pii/S1386418111000176>.
- Orleans Martins and Edilson Paulo. A probabilidade de negociação com informação privilegiada no mercado acionário brasileiro. *Brazilian Review of Finance*, 11:249, 07 2013. doi: 10.12660/rbfin.v11n2.2013.6233.
- Orleans Martins, Edilson Paulo, and Pedro Albuquerque. Negociação com informação privilegiada e retorno das ações na bmf-bovespa. *Revista de Administração de Empresas*, 53:350–374, 08 2013. doi: 10.1590/S0034-75902013000400003.
- Murat Tiniç and Duygu Celik. Infotrad: An r package for estimating the probability of informed trading. *The R Journal*, 10, 12 2017. doi: 10.32614/RJ-2018-013.
- Yuxing Yan and Shaojun Zhang. An improved estimation method and empirical properties of pin. 01 2006.

¹⁶The association between Novo Mercado and lower PIN does not imply that Novo Mercado has any causal effect on the PIN. Perhaps companies that are less susceptible to insider trading self-select into Novo Mercado. Also, trading activity is a confounder here: companies in Novo Mercado have stocks that are traded more frequently than companies outside Novo Mercado, and trading activity is inversely correlated with PIN. The causal effects of Novo Mercado are outside the scope of this paper.