Empirical Market Microstructure Analysis (EMMA)

Lecture 7: Information-based Models

Prof. Dr. Michael Stein
michael.stein@vwl.uni-freiburg.de

Albert-Ludwigs-University of Freiburg

Summer Term 2016
Outline

1. Introduction: Financial Markets and Market Structure
3. Statistical Building Blocks and Econometric Basics
4. Transaction and Trading Models
5. Information-Based Models
6. Inventory Models
7. Limit Order Book Models
8. Price Discovery and Liquidity
9. High-frequency Trading
10. Current Developments
Important Notes

- Lecture 6 and 8 are devoted to models that enable to estimate the transaction costs including the component of inventory costs.

- While Lecture 6 introduced inventory models, Lecture 7 and 8 will be used to add information based models and for a thorough discussion of the Madhavan-Smidt model, which combines information and inventory models.

- Lecture 7 contains the discussion of the PIN model (Probability of Informed Trading) by Easley, Kiefer, O’Hara and Paperman (1996), Easley et al. (1996) in the following, which can be seen as an extension of the Glosten-Milgrom (1985) model and other sequential trade models. It therefore serves as a natural selection for discussion of information-based models before the Madhavan-Smidt model combination in Lecture 8.
Outline

Lecture 7: Information-Based Models

Motivation
Easley et al.: PIN Model Setup
Empirical Setup of the PIN Model
Empirical Results of the PIN Model
Origination and Basis

- We directly follow the Easley et al. study to derive the model as the focus and type of derivation is slightly different to the models discussed so far, making notational adjustments redundant. In addition, De Jong and Rindi's derivation of the model is fairly short.
- The Easley et al. paper is very influential in the area of information based models for market microstructure.
- The PIN model of Easley et al. is developed in order to classify trades as either buy or sell based on estimations regarding the probability that informed trading occurred. A difference to models discussed so far is that there is no assumption that there are no news or that news are incorporated by the public - they can be captured by the random term.
Motivation

- Model summarized:

```
ABSTRACT

This article investigates whether differences in information-based trading can explain observed differences in spreads for active and infrequently traded stocks. Using a new empirical technique, we estimate the risk of information-based trading for a sample of New York Stock Exchange (NYSE) listed stocks. We use the information in trade data to determine how frequently new information occurs, the composition of trading when it does, and the depth of the market for different volume-decile stocks. Our most important empirical result is that the probability of information-based trading is lower for high volume stocks. Using regressions, we provide evidence of the economic importance of information-based trading on spreads.
```

- Infrequently traded stocks pose an interesting surrounding, as the prices are less dynamic and spreads are substantial. This may be either due to inventory costs, as market-makers must maintain positions for lengthy periods of time, or market power by the specialist dealers.
Outline

Lecture 7: Information-Based Models

Motivation
Easley et al.: PIN Model Setup
Empirical Setup of the PIN Model
Empirical Results of the PIN Model
Sequential Trade Model Basis

- Trading is modelled with a mixture of discrete and continuous time components.
- Participants in the market are risk-neutral, competitive market makers and traders which may be uninformed or informed.
- While models discussed so far had the arrival of traders as an assumption that mainly served as model simplifications or necessities when trying to estimate price successions, the PIN model directly models the traders’ arrivals with continuous time methods.
- The prices the market-maker sets through bid and ask are results of the expected value of the asset conditioned on his information.
- A single risky asset is used, which is traded over $i = 1, 2, \ldots, I$ trading days, where intraday time is continuous and denoted $t \in [0, T]$. 
Sequential Trade Model Basis

- Information events occur with probability $\alpha$, where out of those news, good news are occurring with probability $1 - \delta$ and bad news with probability $\delta$. At the end of the day, i.e. after trading, all information is realized.

- Random variables $(V_i)_{i=1}^{I}$ denote the value of the asset at the end of the day. There is no assumption needed on autocorrelation of the prices.

- Values conditional on good (bad) news are $\overline{V_i}$ ($\underline{V_i}$). No information leads to $V_i^*$, with $\overline{V_i} > V_i^* > \underline{V_i}$.

- Easley et al. assume that all informed traders are risk neutral and competitive: Good signals lead to buying of the asset and bad signals lead to selling.
Sequential Trade Model Basis

- The model setup leads to the following tree of the sequential trade model:

Figure 1. Tree diagram of the trading process. This figure gives the structure of the trading process, where \( \alpha \) is the probability of an information event, \( \delta \) is the probability of a low signal, \( \mu \) is the rate of informed trade arrival, and \( \epsilon \) is the rate of uninformed buy and sell trade arrivals. Nodes to the left of the dotted line occur once per day.
Trading and Price Building

- The market-maker does not know which information event happened, and forms his prior beliefs for no, bad and good news \((n, b \text{ and } g)\) as: \(P(t) = (P_n(t), P_b(t), P_g(t))\).

  At \(t = 0\), his prior is accordingly: \(P(0) = (1 - \alpha, \alpha \delta, \alpha (1 - \delta))\)

- He is forming prices dependent on order arrivals and conditional on the process before order arrivals:

  Posterior probability: no news and sell order \(S_t\) arrives at
  \[t: P_n(t \mid S_t) = \frac{P_n(t)\varepsilon}{\varepsilon + P_b(t)\mu}\]

  Posterior probability: bad news and sell order \(S_t\) arrives at
  \[t: P_b(t \mid S_t) = \frac{P_b(t)(\varepsilon + \mu)}{\varepsilon + P_b(t)\mu}\]

  Posterior probability: good news and sell order \(S_t\) arrives at
  \[t: P_g(t \mid S_t) = \frac{P_g(t)\varepsilon}{\varepsilon + P_b(t)\mu}\]
Trading and Price Building

• The bid price $b(t)$ at time $t$ (which is a zero-profit bid price as market makers face competition) is the market maker’s expected value of the asset conditional on the history prior to $t$ and on $S_t$:
  
  $$b(t) = \frac{P_n(t)\varepsilon V^*_i + P_b(t)(\varepsilon + \mu)V_i + P_g(t)\varepsilon \overline{V}_i}{\varepsilon + P_b(t)\mu}$$

• And the ask price accordingly is:
  
  $$a(t) = \frac{P_n(t)\varepsilon V^*_i + P_b(t)\varepsilon V_i + P_g(t)(\varepsilon + \mu)\overline{V}_i}{\varepsilon + P_g(t)\mu}$$

• Note that the expected value of the asset conditional on the history of trade before $t$ is:
  
  $$E[V_i \mid t] = P_n(t)V^*_i + P_b(t)V_i + P_g(t)\overline{V}_i$$

• Combining above equations:
  
  $$b(t) = E[V_i \mid t] - \frac{\mu P_b(t)}{\varepsilon + \mu P_b(t)} (E[V_i \mid t] - V_i)$$
  
  $$a(t) = E[V_i \mid t] + \frac{\mu P_g(t)}{\varepsilon + \mu P_g(t)} (\overline{V}_i - E[V_i \mid t])$$
Trading and Price Building

- From the above derivations, it can be seen that the arrival of information and the probabilities assigned to it, drive the prices the market makers set.
- Easley et al. provide a most intuitive interpretation of the equations:

These equations demonstrate the explicit role played by arrivals of informed and uninformed traders in affecting trading prices. If there are no informed traders ($\mu = 0$), then trade carries no information, and so the bid and ask are both equal to the prior expected value of the asset. Alternatively, if there are no uninformed traders ($\epsilon = 0$), then $b(t) = \bar{V}_i$ and $a(t) = \bar{V}_i$ for all $t$. At these prices no informed traders will trade either, and the market, in effect, shuts down. Generally, both informed and uninformed traders will be in the market, and so the bid is below $\mathbb{E}[V_i|t]$ and the ask is above $\mathbb{E}[V_i|t]$. This spread results from the market maker setting prices to protect himself from losses to informed traders.
Trading and Price Building

• Given the definition of the spread $\sum(t) = a(t) - b(t)$, one obtains:
  $$\sum(t) = E[V_i | t] + \frac{\mu P_g(t)}{\varepsilon + \mu P_g(t)} (\overline{V_i} - E[V_i | t]) - E[V_i | t] + \frac{\mu P_b(t)}{\varepsilon + \mu P_b(t)} (E[V_i | t] - V_i)$$

• Accordingly, the spread at time $t$ is the probability that a buy is information-based multiplied with the expected loss against an informed buyer, plus the analogous term for a sell order by an informed trader.

• The sum of these probabilities is finally representing the probability that a trade is information-based:
  $$PIN(t) = \frac{\mu(1-P_n(t))}{\mu(1-P_n(t)) + 2\varepsilon}$$

• Intuition is again best provided by Easley et al.:
Trading and Price Building

- The first spread from the opening quotes is especially intuitive for the case that good and bad information events are equally probable, i.e. when $\delta = (1 - \delta)$:
  $$\sum(0) = \frac{\alpha \mu}{\alpha \mu + 2 \varepsilon} [V_i - V_i]$$

- Intuition is again best provided by Easley et al:

The first term in this equation is the probability that the first trade of the day is information-based. This risk of trading with an informed trader is clearly a crucial factor influencing the size of spreads. If this probability differs between stocks, then our model predicts how initial spreads will differ, and this provides a way to test for information-based differences in spreads.

If, like the market maker, we knew the parameters of the problem, $\theta = (\alpha, \delta, \varepsilon, \mu)$, and observed the order arrival process, then we could compute the stochastic process of bids and asks. This would allow us to directly examine the effect of information on spreads. Although we can observe the order arrival process, we do not know the parameters. These parameters can be estimated, however, from the data on order arrivals. It is to this problem that we now turn.
Outline

Lecture 7: Information-Based Models

Motivation
Easley et al.: PIN Model Setup
Empirical Setup of the PIN Model
Empirical Results of the PIN Model
Estimation Framework

- While the arrival rates are modelled using a Poisson process, the estimation is done for the parameters $\alpha$, $\delta$, $\varepsilon$, and $\mu$.
- The crucial problem is the fact that while buy and sell orders may be observed, we may not observe information events and which investors are uninformed and which are not.
- This is solved by estimating each day, with the assumption being that good events increase buy transactions and bad events increase sell transactions. Days where there is (almost) no trading may be the result of no information events/ no informed investors. Technically, this is accomplished by mixing the three Poisson processes.
- For a day where bad news happened, sell orders arrive with probability $(\mu + \varepsilon)$ for informed and uninformed traders and buy orders by uninformed market participants arrive with probability $\varepsilon$. 
Estimation Framework

- Given the assumption of continuous Poisson processes, we obtain the following likelihood functions when observing $B$ buys and $S$ sells on any day:
  
  For a bad news day:
  \[ e^{-\varepsilon T} \frac{(\varepsilon T)^B}{B!} e^{-(\mu+\varepsilon) T} \frac{[(\mu+\varepsilon) T]^S}{S!} \]

  For a no news day:
  \[ e^{-\varepsilon T} \frac{(\varepsilon T)^B}{B!} e^{-\varepsilon T} \frac{[(\varepsilon T)^S}{S!} \]

  For a good news day:
  \[ e^{-(\mu+\varepsilon) T} \frac{[(\mu+\varepsilon) T]^B}{B!} e^{-(\mu+\varepsilon) T} \frac{[(\mu+\varepsilon) T]^S}{S!} \]

- Accordingly, the likelihood function using the probabilities of the different news happening, yields the following:
  \[ L((B, S) | \theta) = (1 - \alpha) \cdot e^{-\varepsilon T} \frac{(\varepsilon T)^B}{B!} e^{-\varepsilon T} \frac{[(\varepsilon T)^S}{S!} + \alpha \delta \cdot e^{-\varepsilon T} \frac{(\varepsilon T)^B}{B!} e^{-(\mu+\varepsilon) T} \frac{[(\mu+\varepsilon) T]^S}{S!} + \alpha (1 - \delta) \cdot e^{-(\mu+\varepsilon) T} \frac{[(\mu+\varepsilon) T]^B}{B!} e^{-\varepsilon T} \frac{[(\varepsilon T)^S}{S!} \]

Estimation Framework

- As the maximum-likelihood estimator will be either 0 or 1 for $\alpha$ and $\delta$ due to information occurrence being restricted to occurring once a day, this enables to estimate trader selection probabilities using intra-day analysis.

- Inter-day analysis is informative on the information event parameters, with the product of the daily likelihoods representing the probability of observing the data $M = (B_i, S_i)_{i=1}^l$ for $l$ days.

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

- The estimation procedure is less-clear cut as for other models discussed so far, because the tests are carried out using a structural approach. This means that the estimated probabilities need to be statistically assessed whether they indeed represent differences regarding trading in stocks, and because the estimated values need to be tested on whether they indeed stem from information-based trading.
Data and Technicality

- Easley et al. use a sample of NYSE traded common stocks. The randomly composed sample is constructed using ranking by price and volume of trading.
- As in other models discussed so far, the estimation horizon is kept short with 60 trading days to most probably eliminate non-stationarity of the prices.
- 90 stocks were used finally after the ranking and matching procedures.
- Very quick successive trades within seconds were treated as one.
- Buy and sell classification was done by separating trades above and below the midpoint. Trades at the midpoint get related to the trade before, so-called “tick-test”.
- Easley et al. restrict the parameters $\alpha$ and $\delta$ to be in either $(0, 1)$ using a logit transformation, and $\varepsilon$ and $\mu$ to $(0, \infty)$ by a logarithmic transformation.
- A quadratic hill-climbing algorithm is used for the estimation.
Outline

Lecture 7: Information-Based Models

Motivation
Easley et al.: PIN Model Setup
Empirical Setup of the PIN Model
Empirical Results of the PIN Model
Empirical Results

- Results for the parameter estimates and the PIN:

![Table 1: Summary Parameter Estimate Statistics by Decile](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>First Decile</th>
<th>Fifth Decile</th>
<th>Eighth Decile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number in Sample</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean</td>
<td>0.131970</td>
<td>0.030148</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.104864</td>
<td>0.027596</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
<td>0.079831</td>
<td>0.013238</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Mean</td>
<td>0.175742</td>
<td>0.023970</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.136797</td>
<td>0.022917</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
<td>0.141192</td>
<td>0.013158</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Mean</td>
<td>0.500294</td>
<td>0.433952</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.477761</td>
<td>0.448613</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
<td>0.141192</td>
<td>0.170253</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Mean</td>
<td>0.349078</td>
<td>0.444393</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.360357</td>
<td>0.418164</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
<td>0.227188</td>
<td>0.238763</td>
</tr>
<tr>
<td>Prob(Inf)</td>
<td>Mean</td>
<td>0.163919</td>
<td>0.207788</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.154193</td>
<td>0.205858</td>
</tr>
<tr>
<td></td>
<td>Std. dev.</td>
<td>0.043794</td>
<td>0.064794</td>
</tr>
</tbody>
</table>
Empirical Results

- Results on whether the groups ordered by trading volumes differ significantly, done via population/distribution tests:

Table II
Nonparametric Tests

The Kruskal-Wallis statistic is used to test the null hypothesis that parameter values for all three volume samples are drawn from identical populations versus the alternative hypothesis that at least one of the populations tends to furnish greater observed values than other populations. The parameter $\mu$ is the arrival rate of informed traders, $\epsilon$ is the arrival rate of uninformed traders, $\alpha$ is the probability of an information event, and $\delta$ is the probability that new information is bad news. The parameter Prob(Inf) is a composite variable measuring the probability of information-based trade.

Panel A: Kruskal-Wallis Tests on Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>66.279</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>69.859</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>10.853</td>
</tr>
<tr>
<td>$\delta$</td>
<td>4.236</td>
</tr>
<tr>
<td>Prob(Inf)</td>
<td>8.027</td>
</tr>
</tbody>
</table>

Critical value for $\alpha = 0.05$ is 5.991.

The Mann-Whitney statistic is used to test the null hypothesis that two samples are drawn from identical populations against the alternative that one population tends to yield higher values. The parameter $\mu$ is the arrival rate of informed traders, $\epsilon$ is the arrival rate of uninformed traders, $\alpha$ is the probability of an information event, and $\delta$ is the probability that new information is bad news. The parameter Prob(Inf) is a composite variable measuring the probability of information-based trade.

Panel B: Mann-Whitney Tests on Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>1 to 5</th>
<th>1 to 8</th>
<th>5 to 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>6.402</td>
<td>6.623</td>
<td>4.489</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>6.505</td>
<td>6.603</td>
<td>5.071</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>4.190</td>
<td>3.336</td>
<td>1.789</td>
</tr>
<tr>
<td>$\delta$</td>
<td>1.508</td>
<td>-1.937</td>
<td>-0.547</td>
</tr>
<tr>
<td>Prob(Inf)</td>
<td>-2.883</td>
<td>-1.952</td>
<td>0.192</td>
</tr>
</tbody>
</table>

The test statistic is normally distributed and the critical value for $\alpha = 0.05$ is $1.6449$. 
Parameter Implications

- The first table shows that there are considerable differences between the groups ordered by trading activity, with the probability of information events being smaller for less active stocks than for stocks with high trading volumes. Therefore, the results for the probability of an information event are in line with the model’s theoretical predictions.

- The second table shows that the hypothesis of equal distributions of estimates of probability events for the three ordered groups is strongly rejected, what is in line with the expectations from the model.

- Both the very similar mean results and the population tests that do not reject the null indicate that there is no difference between the groups regarding the probability of bad news. This is reasonable from theoretical considerations as well.
Parameter Implications

- For the arrival rates of informed and uninformed traders, the results from parameter estimates and population tests again are in line with theoretical expectations, as higher volume stocks have higher arrival rates of both uninformed and informed traders.

- While spreads should be smaller for higher volume stocks, the appearance of events and informed traders works in the opposite direction. The “composite measure PIN” is therefore needed to see how the effects are.

- It can be seen from the first table that the probability of informed trading is lowest for the most active stocks, and highest for the least active stocks. This indicates that the PIN is inversely related to volume, other than the probability of information events and the arrival rates of both types of traders.
PIN Results

- However, the difference between decile 5 and 8 is less clear than for other parameters with differences in parameters, making further assessment necessary:

![Graph showing cumulative probability distribution](image-url)

*Figure 2. Cumulative distribution of the probability of information-based trades by volume decile.* This graph shows the cumulative probability distribution of the probability that a trade comes from an informed trader, \( P(\text{Informed} | \text{Trade}) \), for each of the three 1990 volume deciles in our sample.
PIN Implications

- The result of an inverse relationship between PIN and the trading volume does show up between the 1st decile and the other 2, while the 5th and 8th decile even cross sometimes.
- As the 1st, 5th and 8th decile represent high, medium and low volume groups, the implication is that PIN is smaller for very actively traded stocks, compared to others. This should lead to lower spreads for actively traded stocks compared to others, which should not have large spread differences among each other.
PIN Implications

- Again the curves of the deciles reveal more information regarding the differences and similarities:
Spatials and Informed Trading

- While the table above is in line with what should be expected for the spreads, the interesting finding is that the relation in the spreads mirrors the findings of the PIN results, with no large differences between the 5th and 8th decile, but between each of those and the 1st decile.

- The economic significance of the PIN estimates is tested by Easley et al. as well: Using the opening spread being defined as \( \sum = [V_i - V_j] \cdot PI \) and assuming that the price range is a linear function of the stock price \( V \), the spread is expressed as \( \sum = \beta_1 \cdot V \cdot PI \) and therefore a testable equation for spreads and PIN is:
  \[
  \sum = \beta_0 + \beta_1 \cdot V \cdot PI + \beta_2 \cdot Vol + \eta
  \]

- The elegance of this equation is best understood when considering that the constant will capture all pricing effects that are not attributed to the PIN, and that the volume term is controlling for the size effect, leaving the “pure” PIN effect in the relation intact.
Spatrows and Informed Trading

- The regression results strongly confirm the previous findings:

<table>
<thead>
<tr>
<th>General Model</th>
<th>Restriction to $\beta_1 = 0$</th>
<th>Restriction to $\beta_2 = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.2114</td>
<td>0.2885</td>
</tr>
<tr>
<td></td>
<td>(20.453)</td>
<td>(31.954)</td>
</tr>
<tr>
<td>$V \cdot PI$</td>
<td>0.0193</td>
<td>0.0178</td>
</tr>
<tr>
<td></td>
<td>(9.463)</td>
<td>(7.982)</td>
</tr>
<tr>
<td>$Vol$</td>
<td>$-1.035E-11$</td>
<td>$-6.879E-12$</td>
</tr>
<tr>
<td></td>
<td>($-4.572$)</td>
<td>($-2.175$)</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>0.5216</td>
<td>0.0402</td>
</tr>
<tr>
<td>F-Value</td>
<td>49.518</td>
<td>4.730</td>
</tr>
</tbody>
</table>
Interpretations

- The empirical results show a similar behaviour of PIN and spread differences with respect to deciles, and this in both statistically and economically meaningful dimensions.
- Less active stocks appear to be more exposed to informed trading, and the larger spread of those appears to be a result of this, rather of market power or pure trading volume considerations.
- The finding of higher PIN in less frequently traded stocks and the price effect from such asymmetric information effects is in line with other empirical findings.
- Interestingly, the higher arrival rates of informed investors and the higher probability of news events are more than offset by the higher rate of uninformed traders for large volume stocks.
References

F. de Jong and B. Rindi


- In addition, the interested reader may review literature contained in the reference list therein and the papers discussed.