Preliminary Project Update

Nicholas Head

January 11, 2015

Problem Background

VPIN is an extension of the PIN model developed by O'Hara, Lopez and Easley (hereafter known as ELO) used to identify the probability of information-based trading[1]. That is, if one is a non-informed participant such as a market-maker, what is the probability of dealing with an informed trader, and hence becoming subject to adverse selection. This is a critical metric for market makers as it enables them to set appropriate bid-ask prices such that the spread compensates them for the probability of dealing with an insider.

[1] Easley, D., López De Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies, 25, 1457–1493. doi:10.1093/rfs/hhs053

For this reason PIN has been used as a metric to measure the extent of so-called order flow toxicity. If the order flow becomes too toxic, market makers are forced out of the market. As they withdraw, liquidity disappears, which increases even more the concentration of toxic flow in the overall volume, which then triggers a feedback mechanism that forces even more market makers out. This cascading effect has caused liquidity-induced crashes in the past, the 2010 Flash Crash being one (major) example of it[2]. One hour before the flash crash, order flow toxicity was at historically high levels relative to recent history. ELO claim that using the VPIN metric, this crash could have been predicted one hour before it actually happened.

[2] https://en.wikipedia.org/wiki/2010_Flash_Crash

The theoretical underpinning for VPIN is based off inferring toxicity from trade imbalances using so-called volume synchronised time bars. This necessitates using a trade classification algorithm to identify trades as either buys or sells. The procedure has been controversial however with several papers by Andersen and Bondarenko refuting the claims made by ELO.

VPIN's theory is consistent with the anecdotal evidence reported by a joint SEC-CFTC study on the events of May 6, 2010. Given the relevance of these findings, the S.E.C. requested an independent study to be carried out by the Lawrence Berkeley National Laboratory. This Government laboratory concluded:[3]

> This [VPIN] is the strongest early warning signal known to us at this time.

[3] Bethel, W., Leinweber, D., Ruebel, O., & Wu, K. (2011). Federal Market Information Technology in the Post Flash Crash Era: Roles for Supercomputing. SSRN Electronic Journal. doi:10.2139/ssrn.1939522

The work completed to date has been primarily in the areas of data sourcing, calculating simulated PIN and VPIN data, and additionally a large amount of literature review. This report is the result of the work completed so far.

Plan for Remaining Work

Simulation Study

The first task is to identify whether a Hidden Markov Model will execute the parameter estimation correctly. Specifically I will seek to identify whether an HMM can correctly identify the hidden intraday market states and from there conclude whether VPIN is an accurate measure of those states. To this end I will perform a simulation analysis where the hidden states are known beforehand (in terms of the sequential trade model parameters)

The simple sequential trade model is as follows. Denote a security's price as S. Its present value is S_1. Once a certain amount of new information has been incorporated into the price, S will be either S_B (bad news) or S_G (good news). There is a probability α that new information will arrive within the time-frame of the analysis, and a probability δ that the news will be bad (i.e., $1 - \delta$ that the news will be good). Traders are classified into two types: So-called noise, or liquidity traders, are those with no information based reason to trade. Their arrivals are measured as a poisson process with rate ϵ. Information-based traders on the other hand are those that are trading based off some private information that has not been priced into the asset. They are also modelled as a poisson process, with arrival rate μ.

If an information event arrives and it is a 'good news' event, both uninformed and informed traders buy, while only uninformed traders sell. Conversely if a 'bad news' information event occurs only uninformed traders buy, while both informed and uninformed traders sell. If no information event occurs then only uninformed traders are in the market. The model is visually explained the in the decision tree in Figure 1.

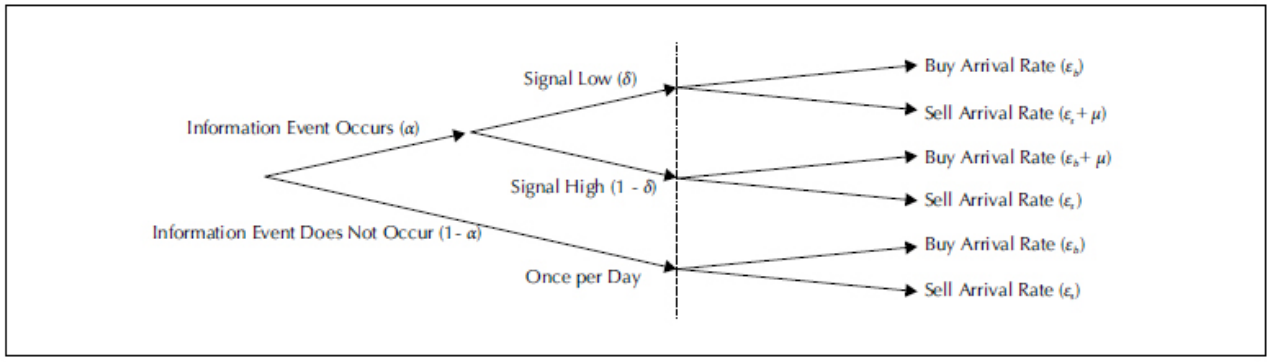

Figure 1 *Tree Diagram of the Trading Process*

Source: Adapted from Easley, Hvidkjaer, and O'Hara (2002).

The original PIN model requires the estimation of four non-observable parameters, namely α, δ, μ, and $epsilon$. This was originally done via Maximum likelihood, through the fitting of a mixture of three Poisson distributions. VPIN can in some regards be considered a high-frequency estimate of PIN which takes into account the time-varying nature of the data. Instead of calculating liquidity based off clock-time, VPIN uses what is called volume-time, whereby trades are placed into equally sized buckets each with a uniform amount of volume. The intuition behind this is that in a high-frequency environment, these volume bars will be spaced out widely when there is little information-based trading, and packed tightly when there is a higher amount. The idea is that more relevant is a piece of information, more volume it will attract. Within each of these bars the number of trades that are buys or sells are inferred, and then VPIN is calculated as the degree of imbalance between buys and sells. If we have $\tau = 1, 2, ...n$ volume buckets, within each bucket trades are classified as buys V_τ^B, and sells V_τ^S where $V = V_\tau^B + V_\tau^S$ for each τ. Using this approach, the parameters can be estimated anlytically instead of numerically.

A crucial part of this calculation involves the classification of trades into buys and sells, as traditional order-book data sources do not include the trade direction. Rather than using the Tick-rule, Lee-Ready or other trade classification techniques, ELO propose a new volume classification method called Bulk Volume Classification. This departs from standard trade classification schemes in two ways: First, volume is classified in bulk, and second this methodology classifies part of a bar's volume as buy, and the remainder as sell. Empirical studies have shown Bulk Volume Classification to be more accurate than the Tick-rule, despite of not requiring level-1 tick data (only bars)[4]. Within a volume bucket, the amount of volume classified as buy is

[4] Easley, D., Lopez de Prado, M. M., & O'Hara, M. (2012). Bulk Classification of Trading Activity. SSRN Electronic Journal. doi:10.2139/ssrn.1989555

$$V_\tau^B = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i Z\left(\frac{S_i - S_{i-1}}{\sigma_{\Delta S}}\right) \tag{1}$$

where $t(\tau)$ is the index of the last (volume or time) bar included in bucket τ, V_τ^B is the buy volume (traded against the Ask), V_i is the total volume per bucket, Z is the Standard Normal Distribution, and $\sigma_{\Delta S}$ is the standard deviation of price changes betwenen (volume or time) bars. Because all buckets contain the same amount of volume V,

$$V_\tau^S = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \left(1 - Z\left(\frac{S_i - S_{i-1}}{\sigma_{\Delta S}}\right)\right) = V - V_\tau^B \tag{2}$$

The algorithm to compute the VPIN metric is in Appendix A.

The dynamics of buys and sells are driven by the sequential trade model parameters mentioned previously (α, δ, μ and ϵ). The expected arrival rate of informed trade becomes: $E[V_\tau^B - V_\tau^S] = \alpha\mu(2\delta - 1)$ and the absolute expected values for sufficiently large μ is $E[|V_\tau^B - V_\tau^S|] \approx \alpha\mu$.

The total expected arrival rate is:

$$\frac{1}{n}\sum_{\tau=1}^{n}(V_\tau^B + V_\tau^S) = V =$$

$$\underbrace{\alpha(1-\delta)(\epsilon+\mu+\epsilon)}_{\text{Volume from good news}} + \underbrace{\alpha\delta(\mu+\epsilon+\epsilon)}_{\text{Volume from bad news}} + \underbrace{(1-\alpha)(\epsilon+\epsilon)}_{\text{Volume from no news}} = \alpha\mu + 2\epsilon$$

VPIN is then calculated as:

$$VPIN = \frac{\alpha\mu}{\alpha\mu + 2\epsilon} = \frac{\alpha\mu}{V} = \frac{\sum_{\tau=1}^{n}(V_\tau^S - V_\tau^B)}{(2\delta-1)nV} \approx \frac{\sum_{\tau=1}^{n}|V_\tau^S - V_\tau^B|}{nV}$$

To simulate the volume bars across the course of a hypothetical trading day I generate data according to the following procedure: Generate a Bernouilly random variable with parameter α to determine whether an information event has ocurred within this volume bar. Then generate a second Bernouilly R.V. with parameter δ to determine whether the event was a 'good' or a 'bad' information event. According to the sequential trade model we then simulate the number of buys and sells for the volume bar using Possion R.V.s. This is repeated for the total number of expected volume bars for the trading days, then this hypothetical day can be simulated a large number of times in order to calculate an expected VPIN value. The code to generate Monte Carlo simulations of the volume bars has been outlined in Appendix B.

A further extension of the basic model is one where the arrival rates of informed and uninformed trades are time-varying. Following the results of Easley, Engle, O'Hara and Wu (2008)[5] the arrival rates can be modelled as a bivariate vector autoregressive process where $(\alpha\mu_t, 2\epsilon_t)'$ contains the time-$(t-1)$ forecast of the arrival rates at time t, b_{t-1} and s_{t-1} are observed buy and sell orders at time $t-1$, $\odot$ denotes the Hadamard product, the vector $g = (g_1, g_2)'$ captures the growth rates of the two intensities, ω is a 2×1 parameter vector and ϕ and ψ are 2×2 parameter matrices:

$$\begin{pmatrix}\alpha\mu_t \\ 2\epsilon_t\end{pmatrix} = \omega \odot e^{g(t-1)} + \phi\left[\begin{pmatrix}\alpha\mu_{t-1} \\ 2\epsilon_{t-1}\end{pmatrix} \odot e^{g(t-1)}\right] + \psi\begin{pmatrix}|b_{t-1} - s_{t-1}| \\ b_{t-1} + s_{t-1} - |b_{t-1} - s_{t-1}|\epsilon_t\end{pmatrix}$$

[5] Easley, D., Engle, R. F., O'hara, M., & Wu, L. (2008). Time-varying arrival rates of informed and uninformed trades. Journal of Financial Econometrics, 6, 171–207. doi:10.1093/jjfinec/nbn003

This simulated data will then be modelled as an HMM and the hidden states will be subsequently decoded using the EM algorithm. I shall use the Akaike Information Criterion (AIC) to identify the most appropriate model. The selecte model shall be evaluated be seeing its effectiveness at identifying the most likely hidden state for each volume bucket. This shall be measured as a missclassification rate. It is anticipated that an HMM will be well suited to modelling the time-varying nature of the VPIN time series.

One complication may arise with the number of hudden states the model suggests. That is the number of states may not lend itself to any sort of intuitive interpretation. Previous studies have shown that using the k-means clustering algorithm may assist in grouping related states together[6].

[6] Yin, X., & Zhao, J. (2014). A Hidden Markov Model Approach to Information-Based Trading: Theory and Applications. Retrieved from `http://papers.ssrn.com/abstract=2412321`

Empirical data analysis

Once an appropriate HMM model has been identified and m the optimum number of hidden states has been established, I am then planning on letting this model run on a selection of real order book data. I have obtained access to consolidated order book data from the NYSE TAQ database. Studies have shown that even though TAQ data is unsigned (e.g. buys and sells are not labelled) the Bulk Volume Classifier employed by ELO is as accurate (and sometimes more accurate) than the signed order book feed from the NASDAQ INET platform.[7]

[7] Chakrabarty, B., Pascual, R., & Shkilko, A. (2012). Trade Classification Algorithms : A Horse Race between the Bulk-based and the Tick-based Rules.

Using this real order book data I will extract order book data around known liquidity crashes since 2010. The plan is then to see what sort of relationship there is between the VPIN metric and the evolution of the hidden states over the course of these known market crashes. If a high (e.g. 95th percentile) value of the VPIN CDF correlates closely with any set of hidden states we can then conclude that VPIN is indeed predicting (or describing) some form of abnormal liquidity event.

Further tasks

- To see whether these extreme hidden states are exclusively associated with these liquidity crashes or is there are large number of false positives.
- Does VPIN indeed predict these crashes or is it simply describing them as they happen.

Appendix A: Algorithm to Compute the VPIN Metric

Based off algorithm found here[8].

[8] http://rof.oxfordjournals.org/content/suppl/2014/09/25/rfu041.DC1/Web_appedix.pdf

Data:

1. Time series of transactions of a particular instrument (T_i, P_i, V_i)
   (a) T_i: Time of the trade.
   (b) P_i: Price at which securities were exchanged.
   (c) V_i: Volume exchanged
2. V: Volume size (determined by user of the formula)
3. n: Sample of volume buckets used in the estimation.

Result: Prepare Equal Volume Buckets

1. Sort transactions by time ascending: $T_{i+1} \geq T_i, \forall i$
2. Compute $\Delta P_i, \forall i$
3. Expand the number of observations by repeating each observation ΔP_i as many times as V_i. This generates a total of $I = \sum_i V_i$ observations ΔP_i.
4. Re-index ΔP_i observations, $i = 1, ..., I$
5. Initiate counter: $\tau = 0$
6. **while** $\tau V < I$ **do**
   (a) Add one unit to τ
   (b) $\forall i \in [(\tau - 1)V + 1, \tau V]$, split volume between buy or sell intitiated:
      i. Assign to Vb the number of observations classified as buy: $V_\tau^B = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i Z\left(\frac{S_i - S_{i-1}}{\sigma_{\Delta S}}\right)$
      ii. Assign to Vs the number of observations classified as sell: $V_\tau^S = \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \left(1 - Z\left(\frac{S_i - S_{i-1}}{\sigma_{\Delta S}}\right)\right) = V - V_\tau^B$

   end
7. Set $L = \tau - 1$

Appendix B: Monte-Carlo Simulation of Volume Buckets

```
numSim = 1000
Vpin = numeric(numSim)

V = 50  #size of each volume bucket
n = 100  #number of volume buckets

alpha = 0.5  #probability of information event
delta = 0.4  #probability of bad news event
mu = 10  #arrival rate of informed trader
epsilon = (V - alpha * mu)/2  #arrival rate of uninformed trader

s = 1
while (s < numSim) {

    Vbuy = numeric(n)  #buy volume buckets
    Vsell = numeric(n)  #sell volume buckets

    j = 1

    while (j <= n) {

        u1 = runif(1)
        u2 = runif(1)

        if (u1 < alpha) {
            # we have an information event

            if (u2 < delta) {
                # its a bad news event

                # only uninformed traders buy when there's bad
                # news
                Vbuy[j] = rpois(1, epsilon)

                # both informed and uninformed traders sell
                # when there's bad news
                Vsell[j] = rpois(1, mu + epsilon)
            } else {
                # its a good news event

                # both informed and uninformed traders buy
                # when there's good news
```

```
            Vbuy[j] = rpois(1, mu + epsilon)

            # only uninformed traders sell when there's
            # good news
            Vsell[j] = rpois(1, epsilon)
        }
    } else {
        # no information event

        # uninformed traders buy and sell in equal
        # quantities
        Vbuy[j] = rpois(1, epsilon)
        Vsell[j] = Vbuy[j]
    }

    j = j + 1
}

Vpin[s] = sum(abs(Vbuy - Vsell))/sum(Vbuy +
    Vsell)

s = s + 1
}
meanVpin = sum(Vpin)/numSim
varVpin = sum(Vpin^2)/(numSim - 1) - (sum(Vpin)/(numSim *
    (numSim - 1)))^2
```