Preliminary Project Update

Nicholas Head

January 11, 2015

Problem Background

Outline of VPIN and what it is trying to resolve.

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. That is if one is a non-informed participant such as a market-maker, what is the probability that you are dealing with an informed trader, and hence become 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.

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 triggers a feedback mechanism that forces even more market makers out. This cascading effect has caused hundreds of liquidity-induced crashes in past, the 2010 Flash Crash being one (major) example of it. One hour before the flash crash, order flow toxicity was at historically high levels relative to recent history. [1] ELO claim that using the VPIN metric, this crash could have been predicted one hour before it actually happened.

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:[2]

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

Outline algorithm for computing VPIN (`http://rof.oxfordjournals.org/content/suppl/2014/09/25/rfu041.DC1/Web_appedix.pdf`)

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

[2] 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`

Work to Date

Literature review

Data cleansing and preparation

Plan for Remaining Work

Simulation Study

- The first task is to identify whether the proposed Hidden Markov Model will execute the parameter estimation correctly. Specifically I will seek to identify whether the 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).
  TODO: Flesh this out.
- To cater for the real-time intraday nature of the VPIN metric I will have to extend the sequential trade model to take into account the time-varying nature of the data[3]. TODO: Flesh this out.
- This simulated data will then be modelled as an HMM and the hidden states will be subsequently decoded using the EM algorithm. I antocipate that EM will be an appropriate algorithm to use as it is designed to cater for 'missing data' scenarios. If EM is found to be unsatisfactory, MLE may be used as a fallback.

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 Calssifier employed by ELO is as accurate (and sometimes more accurate) than the signed order book feed from the NASDAQ INET platform.[4]

[3] 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

[4] 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 then are:

- 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.