Abnormal Trading Activity Detection

Franco Ho Ting Lin, Sebastian Jaimungal (UofT), Charlie Frantowski (TMX), Abe Chan (TMX)

Introduction

In the current era of high-frequency trading (HFT), institutional investors are disadvantaged because of their relative high-latency compared to co-located traders. To level the playing field, we seek to understand what features are present in the market when there is normal versus abnormal trading activity. To do so, we use various cluster analysis methods like Gaussian Mixture Models to determine what conditions are better for slower traders. Due to the high dimensionality of the data, we also explore dimensional reduction techniques before clustering (Principal Component Analysis + GMM), and those that simultaneously perform dimensional reduction and clustering (DEC).

Data

- Data is aggregated for all the new orders, cancellations and fills/partial fills of securities
- The order book is reconstructed through time
- There are generally two main order types — Limit Orders(LO) and Market Orders(MO)

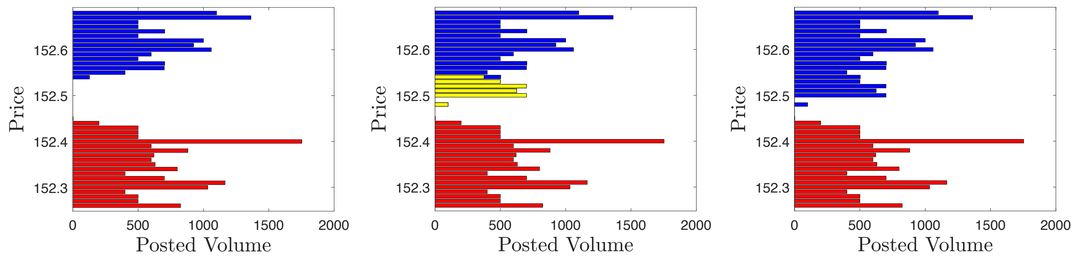

Figure 1. An example of the Limit Order Book. The blue (red) bars represent the volume on the sell (buy) side. The yellow bars indicate which LOs match the incoming MOs.

Abnormal Trading Behaviour

- Spoofing – false or misleading order activity by adding fake supply(demand) and capitalizing on the price deviations
- Quote Stuffing – excessive order activity with intention to interfere with market conditions
- Momentum Ignition – triggers a number of other participants to trade quickly

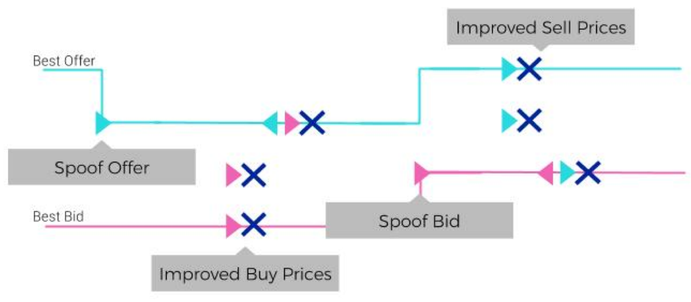

Figure 2. Examples of a Spoofing orders on the Buy side and Sell side

Features

- The feature used are as follows:
  - Volatility
  - Rate of Trade
  - Rate of New Order/ Rate of New Order Touch
  - Rate of Cancelled/ Rate of Cancelled Touch
  - Volume Trade
  - Volume New Order/ Volume New Order Touch
  - Volume Cancelled/ Volume Cancelled Touch

Figure 3. An example of the Rate of Trade and Rate of Change at the Touch over the period of a day.

Cluster Analysis

- Different GMMs were used for different assets
- As we move from high liquidity to low liquidity, it becomes more difficult to pick out distinct clusters

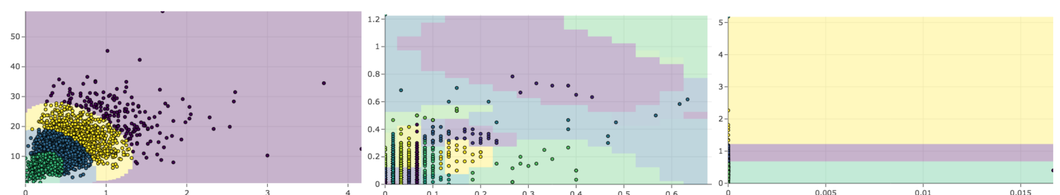

Figure 4. High, medium and low liquidity bins and their cluster boundaries

Dimensionality Reduction + Cluster

- PCA from high dimensional orderbook data to two dimensional latent space
- We then cluster on the latent space

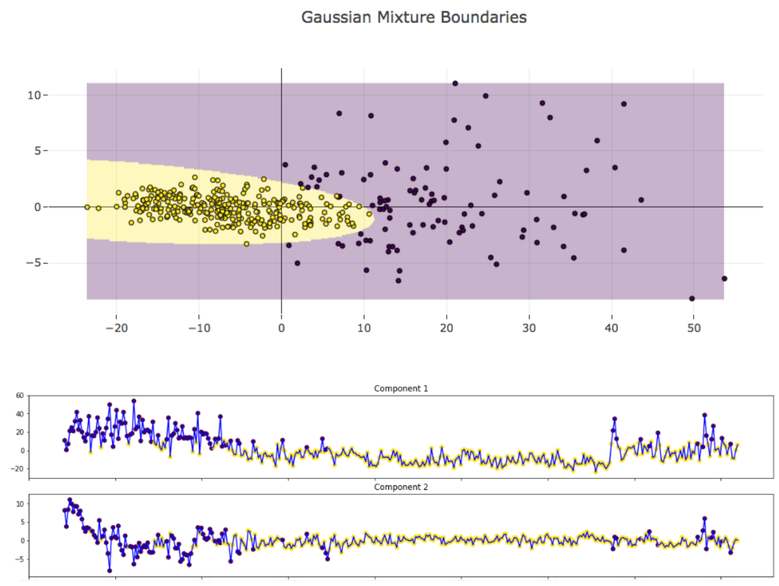

Figure 5. There are two distinct clusters—low/high activity. During the opening and closing times of the market, we are in the high activity cluster; throughout the day, we move into the low activity cluster.

Further Work

- Using an Autoencoder then clustering with a GMM
- Simultaneously performing dimensionality reduction and clustering (DEC [1], DAGMM [2]) has shown promising results for high dimensions

[1] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, JMLR.org, 2016, pp. 478–487.
[2] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, Deep autoencoding gaussian mixture model for unsupervised anomaly detection, in: International Conference on Learning Representations.