MASc Research Weekly Update UTIAS SRG
Anton Rubisov September 14, 2014

Framework

Our goal is to use the dynamics of the Limit Order Book (LOB) as an indicator for high-frequency stock price movement, thus enabling statistical arbitrage. Formally, we will the study limit order book imbalance process, $I(t)$, and the stock price process, $S(t)$, and attempt to establish a stochastic relationship $\dot{S} = f(S, I, t)$.

LOB imbalance, the ratio of limit order volumes between the bid and ask sides, is calculated as $I(t) = \dfrac{V_b(t) - V_a(t)}{V_b(t) + V_a(t)} \in [-1, 1]$.

Modeling $I(t)$: Continuous Time Markov Chain

Instead of modeling imbalance directly, an alternative approach is to discretize imbalance into subintervals (called bins), and model a stochastic process that tracks which bin $I(t)$ falls into. A naive model that can be employed is a continuous-time Markov chain (CTMC).

Let $Z(t)$ be a CTMC taking values in $\{1, \ldots, K\}$, and having infinitessimal generator matrix $\boldsymbol{G}$.[1] Conditional on being in some regime k, the arrival of buy and sell market orders follow independent Poisson processes with intensities $\lambda_k^{\pm}$ (and are hence Markov-modulated Poisson processes), where λ_k^{+} (λ_k^{-}) is the rate of arrivals of market sells (buys).

Given a set of observations of buy/sell market orders and regime switches, we previously derived a maximum likelihood estimation (MLE) for both the entries of $\boldsymbol{G}$ and the values $\lambda_k^{\pm}$. Where $\boldsymbol{G} = \{q_{ij}\} \in \mathbb{R}^{K \times K}$, the q_{ij} represent the transition rates from bin i to j for $i \neq j$, and $q_{ii} = -\sum_{j \neq i} q_i j$ such that the rows sum to 0. We found that:

$$\hat{q}_{ij} = \frac{N_{ij}(T)}{H_i(T)}$$

where

$$N_i j(T) \equiv \text{number of transitions from bin } i \text{ to } j \text{ up to time } T$$
$$H_i(T) \equiv \text{holding time in bin } i \text{ up to time } T$$

Similarly, for the Poisson process intensities $\lambda_k^{\pm}$, we found:

$$\hat{\lambda}_k^{\pm} = \frac{N_k^{\pm}(T)}{H_k(T)}$$

[1] Define the terms $P_{ij}(t) = P\{Z(t) = j | Z(0) = i\}$. Then the matrices $\boldsymbol{P}(t) = \{P_{ij}(t)\}$ and $\boldsymbol{G}$ satisfy $\dot{\boldsymbol{P}}(t) = \boldsymbol{G} \cdot \boldsymbol{P}(t)$, and hence $\boldsymbol{P}(t) = e^{\boldsymbol{G}t}$

where

$$\begin{aligned} N_ij(T) &\equiv \text{number of market orders in bin } k \text{ up to time } T \\ H_i(T) &\equiv \text{holding time in bin } k \text{ up to time } T \end{aligned}$$

Calibrating a CTMC

We estimated parameters for a CTMC on a day's worth of LOB data. Using these parameters, we generated sample paths of the imbalance bins as well as arrival of market orders, and re-estimated parameters along the sample paths. By doing this for 10,000 paths we obtained histograms for the parameters (the individual entires of $\boldsymbol{G}$ as well as the intensities $\lambda_k^{\pm}$).

Using data for `ORCL` from 2013-05-15, averaging imbalances over a 100ms window, and taking the number of bins $K = 3$, we obtained the following mean values for the parameters:

$$\boldsymbol{G} = \begin{pmatrix} -0.112 & 0.098 & 0.0122 \\ 0.099 & -0.21 & 0.111 \\ 0.0115 & 0.112 & -0.1235 \end{pmatrix}$$

$$\boldsymbol{\lambda} = \begin{array}{c} \\ + \\ - \end{array} \begin{array}{ccc} k=1 & k=2 & k=3 \\ \multicolumn{3}{c}{\begin{pmatrix} 0.121 & 0.081 & 0.048 \\ 0.0263 & 0.062 & 0.153 \end{pmatrix}} \end{array}$$

Modeling $I(t)$: Hidden Markov Model

Knowing that trading intensity varies over the course of a day, and suspecting possible other sources of double-stochasticity, we tried to fit a Hidden Markov Model (HMM) to the imbalance process $I(t)$. An HMM is specified with the following parameters:

- Underlying hidden states $S = \{S_1, \ldots, S_N\}$
- Observation symbols $V = \{v_1, \ldots, v_M\}$ (a discrete alphabet)
- State transition probabilities $\boldsymbol{A} = \{a_{ij}\}$, where $a_{ij} = P\{q_{t+1} = S_j \mid q_t = S_i\}$
- Observation symbol probability distribution $\boldsymbol{B} = \{b_j(k)\}$,
  where $b_j(k) = P\{v_k \text{ at time } t \mid q_t = S_j\}$
- Initial state distribution $\boldsymbol{\pi} = \{\pi_i\}$

A short-hand for the parameter space is $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$.

Fitting an HMM to an observation sequence $O = O_1 O_2 \ldots O_T$ can be achieved via the Baum-Welch algorithm (or equivalently the Expectation-Modification algorithm), which seeks to adjust λ to maximize $P(O \mid \lambda)$. There is no analytical solution for an optimal λ, and in fact no way of estimating λ given a finite O. However, the Baum-Welch algorithm allows us to locally optimize $P(O \mid \lambda)$.

Calibrating an HMM

We ran the Baum-Welch algorithm (using the pre-packaged `hmmtrain` function in MATLAB) on the same data that was used for the CTMC calibration. Using either 2 or 3 hidden states, and between 3 and 5 imbalance bins, the HMM training algorithm would converge in ~13 steps, and yield a matrix $\boldsymbol{A}$ (hidden state transitions) with values ~1 on the diagonals and ~0 on the off-diagonals.

We discovered that increasing the imbalance averaging time (which affects the observation data fed to the training algorithm) has reducing the diagonal values away from 1. As the averaging time approached 60s, the diagonal values converged to about 0.85.

Next Steps

1. Run cross-validation on the old CTMC imbalance model, also varying the averaging time.

2. Check for a unit root in the imbalance time series using the augmented Dickey-Fuller test, after transforming the data using the logit function.

3. Instead of running a HMM where the hidden state informs the observable imbalance, try having the hidden state affect the transition matrix between imbalance states.

4. Consider a CTMC where the state is actually the pair (I_{k-1}, I_k), with a $k^2 \times k^2$ transition matrix. Cross-validate and compare with regular CTMC.

5. Same as above but with HMM.