Limit Order Book Dynamics

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 $\dot{S} = f(S, I, t)$. We will then attempt to derive an optimal trading strategy based on the observed relationship.

Recap Next Steps

1. Validate previous CTMC cross-validation results. In particular, to calculate the invariant distribution use $\boldsymbol{A} = e^{\Delta t \boldsymbol{G} n}$, where Δt is the size of the timestep and n is the number of steps to the invariant.

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. 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.

4. Same as above but with HMM.

5. Calibrate HMM for the joint distribution $(I_k, \Delta S_k)$.

6. Extra Reading: Bellman Equations, MDP, Partially Observable MDP

Cross-validation of CTMC

This is following up on the cross-validation results from last time. In those results, in order to obtain the invariant distribution for the Markov chain, we calculated a transition probability matrix $\boldsymbol{A}$ for the embedded discrete-time Markov chain and took matrix powers $\boldsymbol{A}^n$ until it converged, and then observed the average number of timesteps that it took to see n transitions in the data.

In these results, we instead use the relationship $\dot{\boldsymbol{P}}(t) = \boldsymbol{P}(t)\boldsymbol{G} \Rightarrow \boldsymbol{P}(t) = e^{t\boldsymbol{G}}$. Thus we calculate the invariant distribution using the averaging time Δt and the number of such timesteps n and observe when $e^{\Delta t \boldsymbol{G} n}$ converges. This value n immediately tells us the timewindow size to remove for cross-validation.

num bins averaging time	stationary n	Timewindow size	err	$\boldsymbol{Err}$
3 bins, 100ms	172	13629 steps (5.8% of series)	0.020345	50% - 464%
3 bins, 500ms	150	2982 steps (6.4% of series)	0.012794	37% - 331%
3 bins, 1000ms	120	1412 steps (6.0% of series)	0.010105	33% - 264%
3 bins, 2000ms	121	866 steps (7.4% of series)	0.006786	27% - 185%
3 bins, 3000ms	129	715 steps (9.2% of series)	0.005877	24% - 214%
3 bins, 5000ms	114	497 steps (10.5% of series)	0.004026	15% - 154%
3 bins, 10000ms	134	476 steps (20.3% of series)	0.001945	7% - 84%
3 bins, 20000ms	167	492 steps (42% of series)	0.001326	5% - 33%
5 bins, 100ms	53	2509 steps (1.1% of series)	0.140246	96% - 7419%
5 bins, 500ms	46	554 steps (1.18% of series)	0.034655	66% - 1035%
5 bins, 1000ms	40	289 steps (1.24% of series)	0.023179	63% - 911%
5 bins, 2000ms	37	168 steps (1.44% of series)	0.012056	52% - 1441%
5 bins, 3000ms	38	137 steps (1.76% of series)	0.009366	44% - Inf%
5 bins, 5000ms	32	93 steps (1.99% of series)	0.006778	41% - 529%
5 bins, 10000ms	37	83 steps (3.55% of series)	0.003355	28% - Inf%
5 bins, 20000ms	29	56 steps (4.79% of series)	0.002009	23% - Inf%
5 bins, 30000ms	29	53 steps (6.79% of series)	0.001533	21% - Inf%

Table 1: Previous results, convergence threshold 1e-04

num bins averaging time	stationary n	Timewindow size	err	$\boldsymbol{Err}$
3 bins, 100ms	478	47.8s (0.2% of series)	0.356402	644% - 11371%
3 bins, 500ms	144	72s (0.3% of series)	0.087631	236% - 985%
3 bins, 1000ms	89	89s (0.4% of series)	0.050605	150% - 480%
3 bins, 2000ms	57	114s (0.5% of series)	0.032076	122% - 725%
3 bins, 3000ms	45	135s (0.6% of series)	0.023662	98% - 552%
3 bins, 5000ms	35	175s (0.75% of series)	0.014182	70% - 514%
3 bins, 10000ms	29	290s (1.2% of series)	0.007361	52% - 496%
3 bins, 20000ms	22	440s (1.9% of series)	0.004447	43% - 1698%
5 bins, 100ms	546	54.6s (0.2% of series)	0.162690	452% - 6785%
5 bins, 500ms	162	81s (0.3% of series)	0.046204	187% - 2590%
5 bins, 1000ms	100	100s (0.4% of series)	0.029900	136% - 2962%
5 bins, 2000ms	65	130s (0.6% of series)	0.017340	86% - 2141%
5 bins, 3000ms	52	156s (0.7% of series)	0.012505	87% - Inf%
5 bins, 5000ms	42	210s (0.9% of series)	0.008035	66% - 978%
5 bins, 10000ms	31	310s (1.3% of series)	0.004563	45% - Inf%
5 bins, 20000ms	25	500s (2.1% of series)	0.002485	42% - Inf%

Table 2: New results, convergence threshold 1e-05

The large errors seen in the error matrix $\boldsymbol{Err}$ are attributable to the corner elements: in the case of 3 bins, this would be G_{13} and G_{31}. Or, for example, the error matrices for 5 bins at 100ms and at 20000ms looked like:

$$\boldsymbol{Err}_{100ms} = \begin{bmatrix} 6.86 & 8.48 & 5.92 & 9.68 & 11.02 \\ 7.57 & 6.82 & 8.80 & 67.58 & 8.31 \\ 6.33 & 5.08 & 4.52 & 8.55 & 16.79 \\ 14.64 & 54.50 & 8.12 & 6.41 & 7.77 \\ 6.82 & 36.76 & 5.47 & 5.86 & 5.04 \end{bmatrix}$$

$$\boldsymbol{Err}_{20000ms} = \begin{bmatrix} 0.79 & 0.99 & 3.63 & 20.23 & Inf \\ 1.10 & 0.44 & 0.82 & 1.36 & NaN \\ 2.07 & 0.64 & 0.42 & 0.88 & 3.83 \\ 3.64 & 1.66 & 0.85 & 0.57 & 2.81 \\ NaN & Inf & 1.42 & 1.08 & 0.87 \end{bmatrix}$$

2-dimensional CTMC

Next we considered a CTMC for the joint distribution $(I(t), \Delta S(t))$ where $I(t)$ is the bin corresponding to imbalance averaged over the interval $[t - \Delta t_I, t]$, and $\Delta S(t) = \text{sign}(S(t + \Delta t_S) - S(t))$, considered individually for the best bid and best ask prices. For 3 bins, this was encoded into one dimension $Z(t)$ as follows:

$Z(t)$	Bin $I(t)$	$\Delta S(t)$
1	Bin 1	< 0
2	Bin 2	< 0
3	Bin 3	< 0
4	Bin 1	0
5	Bin 2	0
6	Bin 3	0
7	Bin 1	> 0
8	Bin 2	> 0
9	Bin 3	> 0

Generator matrices $\boldsymbol{G}_{bid}$ and $\boldsymbol{G}_{ask}$ were estimated for the resulting timeseries. These were converted to one-step probability matrices $\boldsymbol{P}_{bid}$ and $\boldsymbol{P}_{ask}$ using the formula $\boldsymbol{P} = e\boldsymbol{G}\Delta t$, where Δt is the imbalance averaging period. What this matrix encodes are the conditional one-step transition probabilities - for each entry $\boldsymbol{P}_{ij}$ we have:

$$\begin{aligned} \boldsymbol{P}_{ij} &= \mathbb{P}\left[Z_n \in j \mid Z_{n-1} \in i\right] \\ &= \mathbb{P}\left[(\rho_n, \Delta S_n) \in j \mid (\rho_{n-1}, \Delta S_{n-1}) \in i\right] \end{aligned}$$

The aim is to use these $\boldsymbol{P}$ matrices to compute conditional probabilities of price changes. For example, we can ask: if we are currently in imbalance bin 1, and previous were also in bin 1 and saw a negative price change, what is the probability of again seeing a negative price change?

Since each state $(\rho_n, \Delta S_n) \in j$ is actually comprised of two states, say $\rho_n \in k, \Delta S_n \in m$, we can re-write these entries of $\boldsymbol{P}$ as being:

$$\begin{aligned} &\mathbb{P}\left[\rho_n \in i, \Delta S_n \in j \mid \rho_{n-1} \in k, \Delta S_{n-1} \in m\right] \\ =\, &\mathbb{P}\left[\rho_n \in i, \Delta S_n \in j \mid B\right] \end{aligned}$$

where we're using the shorthand $B = (\rho_{n-1} \in k, \Delta S_{n-1} \in m)$ to represent the states in the previous timestep. Using Bayes' Rule, we can write:

$$\mathbb{P}\left[\Delta S_n \in j \mid B, \rho_n \in i\right] = \frac{\mathbb{P}\left[\rho_n \in i, \Delta S_n \in j \mid B\right]}{\mathbb{P}\left[\rho_n \in i \mid B\right]}$$

The left-hand-side value is exactly the conditional probability in price change that we're interested in finding, the numerator is each individual entry of the one-step probability matrix $\boldsymbol{P}$, and the denominator can be computed as:

$$\mathbb{P}\left[\rho_n \in i \mid B\right] = \sum_j \mathbb{P}\left[\rho_n \in i, \Delta S_n \in j \mid B\right]$$

Using 3 bins, 1000ms imbalance averaging, and 500ms price change, we computed $\boldsymbol{P}_{bid}$:

$$\begin{array}{ll}
 & \qquad\qquad \rho_n = 1 \qquad\qquad\qquad\qquad\qquad\qquad \rho_n = 2 \qquad\qquad\qquad\qquad\qquad\qquad \rho_n = 3 \\
\begin{array}{l} \Delta S_n{<}0\rightarrow \\ \Delta S_n{=}0\rightarrow \\ \Delta S_n{>}0\rightarrow \end{array} &
\left[\begin{array}{ccccccccc|ccccccccc|ccccccccc}
.67 & .05 & .04 & .01 & .03 & .04 & .00 & .05 & .05 & .02 & .50 & .12 & .01 & .00 & .02 & .05 & .01 & .02 & .00 & .00 & .52 & .00 & .01 & .00 & .00 & .00 & .00 \\
.33 & .95 & .96 & .99 & .97 & .96 & .41 & .93 & .95 & .96 & .49 & .87 & .98 & .99 & .97 & .91 & .48 & .96 & .98 & .95 & .47 & .95 & .96 & .93 & .98 & .88 & .34 \\
.00 & .00 & .00 & .00 & .00 & .00 & .58 & .02 & .00 & .02 & .01 & .00 & .01 & .01 & .01 & .05 & .51 & .01 & .02 & .04 & .01 & .05 & .03 & .02 & .02 & .12 & .66
\end{array}\right] \\
 & \quad \Delta S_{n-1} < 0 \qquad\qquad\qquad \Delta S_{n-1} > 0 \qquad\qquad\qquad \Delta S_{n-1} = 0
\end{array}$$