FINM37601:Mathematical Market Microstructure Homework 3

Laurent Lanteigne/12279717

December 4th, 2021

Question 1.

$$\begin{aligned}
\mathbb{E}\left[IMB\right] &= \mathbb{E}_N\left[\mathbb{E}\left[\sum_{i=1}^{N} S_i V_i | N\right]\right] \\
&= \mathbb{E}_N\left[N\mathbb{E}\left[S_i V_i\right]\right] = \mathbb{E}_N\left[N\left(p\mathbb{E}[V_i] - (1-p)\mathbb{E}[V_i]\right)\right] \\
&= \mathbb{E}_N\left[N\left(2p-1\right)\mathbb{E}[V_i]\right] \\
&= \mathbb{E}_N\left[N\beta\left(2p-1\right)\right] \\
&= \beta\lambda\tau\left(2p-1\right) \quad \square
\end{aligned} \tag{1}$$

$$\begin{aligned}
\mathbb{E}\left[(IMB)^2\right] &= \mathbb{E}_N\left[\mathbb{E}\left[\left(\sum_{i=1}^{N} S_i V_i\right)^2 \middle| N\right]\right] \\
&= \mathbb{E}_N\left[\mathbb{E}\left[2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(S_i S_j V_i V_j\right) + \sum_{i=1}^{N}\left(S_i^2 V_i^2\right) \middle| N\right]\right] \\
&= \mathbb{E}_N\left[2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbb{E}\left(S_i S_j V_i V_j | N\right) + \sum_{i=1}^{N}\underbrace{\mathbb{E}\left(S_i^2 | N\right)}_{=1}\mathbb{E}\left(V_i^2 | N\right)\right] \\
&= \mathbb{E}_N\left[2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbb{E}\left(S_i S_j\right)\mathbb{E}(V_i)\mathbb{E}(V_j) + 2N\beta^2\right] \\
&= \mathbb{E}_N\left[2\beta^2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\mathrm{Cov}(S_i, S_j) + \mathbb{E}[S_i]\mathbb{E}[S_j]\right) + 2N\beta^2\right] \\
&= \mathbb{E}_N\left[2\beta^2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\mathrm{Corr}(S_i, S_j)\sigma_{S_i}\sigma_{S_j} + (2p-1)^2\right) + 2N\beta^2\right] \\
&= \mathbb{E}_N\left[2\beta^2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\rho^{|i-j|}p(1-p) + (2p-1)^2\right) + 2N\beta^2\right]
\end{aligned}$$

I will use the fact that, given $|r| < 1$,

$$\sum_{k=0}^{N} r^k = \left(\frac{1-r^{N+1}}{1-r}\right) \rightarrow \sum_{k=1}^{N} r^k = \left(\frac{1-r^{N+1}}{1-r}\right) - 1 = \frac{r\left(1-r^N\right)}{1-r}$$

I will also use that

$$2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} 1 = N(N-1).$$

$$\begin{aligned}
\mathbb{E}\left[(IMB)^2\right] &= \mathbb{E}_N\left[\beta^2 N(N-1)(2p-1)^2 + 2N\beta^2 + 2\beta^2 p(1-p)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\rho^{|i-j|}\right] \\
&= \mathbb{E}_N\left[\beta^2 N(N-1)(2p-1)^2 + 2N\beta^2 + 2\beta^2 p(1-p)\sum_{i=1}^{N-1}\sum_{k=1}^{N-i}\rho^{k}\right] \\
&= \mathbb{E}_N\left[\beta^2 N(N-1)(2p-1)^2 + 2N\beta^2 + 2\beta^2 p(1-p)\sum_{i=1}^{N-1}\frac{\rho(1-\rho^{N-i})}{1-\rho}\right] \\
&= \mathbb{E}_N\left[\beta^2 N(N-1)(2p-1)^2 + 2N\beta^2 + \beta^2\frac{2p\rho(1-p)}{1-\rho}\sum_{i=1}^{N-1}\left(1-\rho^{N-i}\right)\right] \\
&= \mathbb{E}_N\left[\beta^2 N(N-1)(2p-1)^2 + 2N\beta^2 + \beta^2\frac{2p\rho(1-p)}{1-\rho}\left(N-1-\frac{\rho(1-\rho^{N-1})}{1-\rho}\right)\right] \\
&= \mathbb{E}_N\left[\beta^2 N(N-1)(2p-1)^2 + 2N\beta^2 + \beta^2 2p\rho(1-p)\left(\frac{N(1-\rho)-1+\rho^N}{(1-\rho)^2}\right)\right]
\end{aligned}$$

I will use the fact that

$$\mathbb{E}(N) = \lambda\tau, \quad \mathbb{E}(N^2) = \lambda\tau(\lambda\tau+1), \quad \mathbb{E}(\rho^N) = e^{\lambda\tau(\rho-1)}$$

$$\begin{aligned}
\mathbb{E}\left[(IMB)^2\right] &= \beta^2\left(\mathbb{E}[N^2]-\mathbb{E}[N]\right)(2p-1)^2 + \mathbb{E}[N]\beta^2 + \beta^2 2p\rho(1-p)\left(\frac{\mathbb{E}[N](1-\rho)-1+\mathbb{E}[\rho^N]}{(1-\rho)^2}\right) \\
&= \beta^2\left(\lambda^2\tau^2(2p-1)^2 + \lambda\tau + 2p(1-p)\rho\left(\frac{\lambda\tau(1-\rho)-1+e^{\lambda\tau(\rho-1)}}{(1-\rho)^2}\right)\right) \\
&= \beta^2\left(\lambda^2\tau^2(2p-1)^2 + \lambda\tau + \frac{2p(1-p)\rho}{(1-\rho)^2}\left(\lambda\tau(1-\rho)-1+e^{\lambda\tau(\rho-1)}\right)\right)
\end{aligned}$$

Using the fact that

$$\mathrm{Var}\left[IMB\right] = \mathbb{E}\left[IMB^2\right] - \mathbb{E}[IMB]^2$$

$$\begin{aligned}
\mathrm{Var}\left[IMB\right] = &= \beta^2\left(\lambda^2\tau^2(2p-1)^2 + \lambda\tau + \frac{2p(1-p)\rho}{(1-\rho)^2}\left(\lambda\tau(1-\rho)-1+e^{\lambda\tau(\rho-1)}\right)\right) - \beta^2\lambda^2\tau^2(2p-1)^2 \\
&= \beta^2\left(\frac{2p(1-p)\rho}{(1-\rho)^2}\left(\lambda\tau(1-\rho)-1+e^{\lambda\tau(\rho-1)}\right) + \lambda\tau\right) \quad \square \qquad (2)
\end{aligned}$$

Question 2. Let's first discuss the data processing that had to be done in order to extract the values. The first step was to create a "continuous" data. What I've done is concatenating all the trading data set together. I've then created a timestamp using the `date` and `time`. Then to make it continuous, after some data exploration and looking at the trading hours of the Shanghai exchanged, there's a 90minutes window in around lunch-time where the exchange is closed. What I've done is shift down by 90minutes the trades in the second part of the day to make it look like a continuous trading period. Now that each day is continuous, what remains is "gluing" different days where the beginning of tomorrow's trading would start exactly after today's end of trading. On a continuous time frame, there's now almost 3 months of continuous trading for a little over one year of data which makes sense when we consider how many trading hours and trading days in a year.

As for estimating the parameters, what I have done is construct a rolling-window of 5 minutes. For each rolling-window I keep track of how many trades there are into a feature `tradeCount`, I also keep track of the the amount of buy and sell from the `BS` column. With the amount of `B` and `S` in each 5 minutes rolling window, I can compute the probability of a Buy or Sell initiated by assuming a Bernoulli process which is stored in `probBuy`. I also remove the first 5 minute of the dataframe to make sure that the starting values of the rolling basis won't affect the statistics.

Using MLE on the 5min rolling window to estimate λ, the non-biased estimate is actually simply the mean of `tradeCount` divided by 300 to find what's the 1s hazard rate. As for β, the MLE estimate is the sum of all the values in `ntrade` divided by the number of values used to compute the sum. For ρ, I've transformed the `BS` column into (1,0,-1) based on if $B \to 1$, if $S \to -1$ and 0 if data is missing. Then I restrict the analysis only for values in the `BS` columns that are non-zero. On this restricted dataframe, I then compute the correlation of `BS` with a one row shift version of itself. As for p, I'm simply taking the mean of `probBuy`.

Parameters	**Value**
λ	1.24
β	354.27
ρ	0.56
p	0.4927

Question 3. The first step was to compute the `MidPrice1` by take the middle point between `AskPrice1` and `BidPrice1` and then generating a predicted `PredSide` where if `price` $>$ `MidPrice1` then `PredSide` $= 1$, if `price` $<$ `MidPrice1` then `PredSide` $= -1$, and else `PredSide` $= 0$. Then I compared `PredSide` to `BS` for both datasets and gathered some statistics which are below.

	SH60051	**SH601398**
Percentage Same Prediction	79.10%	98.34%
Percentage Same Buy Prediction	79.51%	98.29%
Percentage Same Sell Prediction	78.68%	98.37%
Percentage Prediction 0	0.68%	0.02%
Percentage Prediction 0 is Buy	50.82%	50.41%
Percentage Prediction 0 is Sell	49.18%	49.59%

Our naive version of the **Lee and Ready Algorithm** is a lot more accurate on the second dataset. In both cases, the percentage statistics on Buy vs Sell is mostly symmetric. Interesting to note that on the first set, the spread on accuracy on buy vs sell is around $\approx 1\%$ and less than 0.1% for the second set. The percentage of 0 prediction is much higher in the first than second set. I think this is mostly due to the price of the first asset being larger in the first one so relationship between the tick size and the asset price is larger than in the second set. This relationship allows for more granularity and make it more possible for Buyer initiated to be under the mid and vice-versa.

Question 4. As for the methodology here, similar to Question 1 in terms of building the datasets on top of each other. I compute `MidQuote` by the formula given in the assignment. Using the `MidQuote` I can compute the `5minRet` and `1minRet` on a rolling basis based on the window side. In the same windows, I also keep track of the sum of the product of trade side times volume divided by sample volume inside the variable `5minIMB` and `1minIMB`. I also generate `5minInd` and `1minInd` which are simply indicator values on whether `IMB` is positive or negative.

For each subsample I ran a sequentially least square quadratic programming optimizer that takes two inputs (β, γ) and the objective is to minimize the total MSE definied below.

$$\text{MSE}(\beta, \gamma) = \sum_{i=1}^{N} (\beta \sigma_{Ret} |\text{IMB}|^{\gamma} - \text{Ret})^2$$

where the variable Ret is based on `5minRet` or `1minRet`, the variable IMB from `5minIMB` or `1minIMB`, σ_{Ret} is the according sample standard deviation of the according returns computed. Quick note, while working through the data I had to fragment the dataframe based on returns that were less than 1 and greater than -1 a few outliers made the previous calculation impossible without it. I also think it's fair to assume there is not going to be a 100% within one-tick.

As for the results, they are displayed below.

SH600519		
	1-min Window	
	β	γ
+	0.1736	0.4040
-	-0.0686	0.7834
	5-min Window	
	β	γ
+	0.1006	0.5382
-	-0.0764	0.6664
SH601398		
	1-min Window	
	β	γ
+	0.0104	1.1310
-	-0.0663	0.6456
	5-min Window	
	β	γ
+	0.0268	0.8052
-	-0.0062	1.0914

The first thing I noticed is that for negative imbalance the β's are also negative which makes sense and should be expected, more of a sanity check that the result makes sense. The size of the β's are larger in the first set of data than in the second. The size of β also seem to be larger when only looking at positive imbalance except for the 1 minute window in the second dataset. The γ is larger in the second set than in the first set.