FINM 37601: Mathematical Market Microstructure - Fall 2021

Assignment 3 – Imbalance-Return Relationship at Microstructure Level

Due Date: 11:59pm, Dec. 3, 2021, Chicago Time

Introduction: This assignment contains two parts. In the first part, we propose a theoretical model regarding various stochastic processes related to the arrival of trades and calibrate the model with real trade data later on. In the second part, we calibrate and estimate a market impact model with historical tick data.

The total score for this assignment is 100 points. You need to hand in a report in which you should state clearly your steps of data analysis and reasoning, not just a result. You are encouraged to quote lines of your computer program if that can help you explain yourself (not too many, though). You can also attach your computer programs in your hand-in.

Part One:

Suppose that during time interval $[t, t+\tau)$, the number of trades $N(\tau)$ follows a Poisson process, with the arrival rate of λ:

$$P[N(\tau) = k] = \frac{e^{-\lambda\tau}(\lambda\tau)^k}{k!} \quad k = 0,1,\dots \tag{1}$$

Note that $N(\tau)$ is not a function of time *t.* For each trade, we further assume that its size V_i $(i = 1,2,\dots)$ follows an exponential distribution, with the *pdf* being

$$\mathrm{f}(v,\beta) = \frac{1}{\beta}e^{-\frac{v}{\beta}}, \tag{2}$$

while its sign S_i is assumed to follow a Bernoulli process, *i.e.*

$$P[S_i = 1] = p, P[S_i = -1] = 1 - p. \tag{3}$$

$S_i = 1$ means the trade is buyer-initiated, and seller-initiated otherwise.

In this exercise, we assume that the three random processes mentioned above are mutually independent. We also assume that V_i and V_j are uncorrelated, *i.e.* $Corr(V_i, V_j) = 0, if\ i \neq j$. The intuition behind this assumption is that the current

trade size does not affect the size of any future trade. As for trade sign S_i, we assume that $Corr(S_i, S_j) = \rho^{|i-j|}$ (i.e., the sign correlation function is "time symmetric"). From empirical evidences, the signs of consecutive trades can be highly clustered, especially for liquid stocks as well as in an emerging market, where "herding" phenomenon is highly persistent due to the presence of large amount of retail investors.[1]

Thus, the trade imbalance within the time interval $[t, t+\tau)$ is

$$IMB = \sum_{i=1}^{N} S_i V_i. \quad (4)$$

Question 1: (15 points) At any time t, given the above conditions and assumptions, what is the unconditional expectation and unconditional variance of IMB within the next time interval $[t, t+\tau)$? That is, calculate $E(IMB)\ and\ Var(IMB)$.

Hint for Question 1: The law of total expectation can be used here: $E[IMB] = E_N\left[E[\sum_{i=1}^{N} S_i V_i | N]\right]$, $E[IMB^2] = E_N\left[E[\left(\sum_{i=1}^{N} S_i V_i\right)^2 | N]\right]$, as well as $\left(\sum_{i=1}^{N} S_i V_i\right)^2 = 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(S_i S_j V_i V_j\right) + \sum_{i=1}^{N}\left(S_i^2 V_i^2\right)$. Also, keep in mind that $E(S_i S_j) = Cov(S_i, S_j) + E(S_i)E(S_j)$.

Question 2: (30 points) Using the trade data of SH600519 (see attached data files), calibrate parameters λ, β, p and ρ. For the moment, you can set τ to be 5 minutes, or 300 seconds. The observation windows $[t, t+\tau)$ for different time points t's can be overlapping or non-overlapping. By overlapping observation windows, one can collect more sample data for below analyses.

Question 3: (20 points) In the "trade" files attached for both SH600519 and SH601398, the column with header "BS" indicates the sign of each trade, provided by the exchange and can be quite reliable. That is, if it is a "B", it records the trade as buyer-initiated; otherwise, seller-initiated. For trade data that such trade sign information is not provided (there are many markets where the exchange does not easily provide such information), one can estimate the sign of each trade using a relatively simple version of the so call "**Lee and Ready Algorithm**": for each trade,

---

[1] Note that the assumption of sign correlation in the form of $\rho^{|i-j|}$ may contradict the aforementioned assumption of Bernoulli process for trade signs; nevertheless, we use $\rho^{|i-j|}$ to make the derivation explicitly tractable and to reflect empirical evidence of trade sign clustering effect.

compare the trade price with its "prevailing" quote; if the trade price is above the mid-quote of the prevailing quote, label this trade as a "buyer-initiated" trade that has a sign of +1; if the trade price is below the mid-quote of the prevailing quote, label this trade as a "seller-initiated" trade that has a sign of -1; if the trade price is exactly at the mid-quote, assign 0 to its trade sign. Note that mid-quote is defined here as the simple average of bid1 and ask1 of the order book. Please use the data provided to check that, for both SH600519 and SH601398, what is the accuracy of this simple version of the Lee and Ready algorithm, using the exchange-provided information in the column "BS" as a benchmark. Please use as large sample size as possible from the attached files.

Part Two:

In this part, we will calibrate a simple market impact model using the data in the attached files.

On any given trading day, the price return during the time interval $[t, t+\tau)$ can be calculated as

$$Ret = \frac{MidQuote(t+\tau) - MidQuote(t)}{MidQuote(t)}. \quad (5)$$

For the convenience in handling data, you could also try the following definition which, however, is not recommended, as it's sensitive to bid-ask bounce:

$$Ret = \frac{Price(t+\tau) - Price(t)}{Price(t)}. \quad (6)$$

A third way to calculate return series is to use "weighted mid-quote", in which case the MidQuote(t) is defined as the weighted average of bid1 and ask1 by the corresponding askSize1 and bidSize1:

$$MidQuote(t) = \frac{bid1 * askSize1 + ask1 * bidSize1}{askSize1 + bidSize1}. \quad (7)$$

Using MidQuote(t) defined in this way will make the estimate of return series for large bid-ask spread stocks more meaningful as, for such large bid-ask spread stocks, the mid-quote calculated from simple average of bid1 and ask1 rarely changes in time. Note that the *MidQuote(t)* and *Ret* defined in (5) and (7) must be calculated using the "quote" data files attached.

The trade imbalance within the time interval $[t, t+\tau)$ is calculated as

$$IMB = \sum_{i=1}^{N} S_i \frac{V_i}{\overline{V}}. \quad (8)$$

Here $\overline{V}$ is the first moment of trade sizes in the whole sample. Please note that the difference between equation (8) and equation (4). We will use equation (8) for the definition of imbalance going forward.

The relationship between Ret and IMB can be formulated as

$$Ret = \beta \sigma_{Ret} I_{IMB} |IMB|^{\gamma} \ , \quad (9)$$

where $I_{IMB} = 1 \ if \ IMB > 0 \ and -1, otherwise.$

To capture the asymmetric effect of buys vs. sells, the model is slightly extended as

$$Ret = \beta^{+} \sigma_{Ret} |IMB|^{\gamma+}, if \ IMB > 0, \text{ and } Ret = \beta^{-} \sigma_{Ret} |IMB|^{\gamma-}, if \ IMB < 0. \quad (10)$$

In both (9) and (10), σ_{Ret} is the volatility of the stock; it can be calculated (among other ways) as the standard deviation of the time series of returns defined in (5) and (7); it should be calculated across the whole sample period.

Question 4 (35 points): For both SH600519 and SH601398, calibrate the four parameters: $\beta^{+}, \beta^{-}, \ \gamma^{+}$ and γ^{-} in formulas (10) for two cases: (1) τ = 5 minutes, or 300 seconds; (2) τ = 1 minutes, or 60 seconds. Check if there are significant differences between the two cases and comment on them, if any.

Hint for Question 4: Due to the noisy nature of the data, direct linear regression applied to all observations in (9) or (10) will render a model that has rather low R square. To increase the R square statistic, one can "bin" the data in "buckets" of values of *IMB*. For example, one can form groups of observations by quantiles of *IMB*, then average all the *Ret* values in each group. After that, one can regress on such average values of *Ret* on the quantile values of *IMB* to obtain a model.

Note on the Data:

Note1:

The data covers more than one year's tick-by-tick trade data and quote data of two stocks from the Shanghai Stock Exchange that are key members of the CSI300 index, a major equity index for the local market. They're SH601398 (ICBC), and SH600519 (Kweichow Moutai). The price of SH601398 has relatively small intraday variations, while the price of Moutai stock can exhibit high fluctuations. You may get quite different empirical results for these two stocks.

Note2:

As mentioned earlier, when calculating high-frequency returns for ICBC, it will be best if you use mid-quote. However, there're also several definitions for mid-quote. The most common one is $(Bid1Price + Ask1Price)/2$. But this is not recommended for ICBC because it may stay fairly constant throughout the day. The following volume weighted mid-quote definition is a better choice: $MidQuote = \frac{Bid1Price*Ask1Size+Ask1Price*Bid1Size}{Bid1Size+Ask1Size}$.

End of assignment 3.