University of Cape Town

STA5091Z

Data-analysis for High-Frequency Trading

Lab Exam

Hawkes Process Estimation and Simulation

Authors:
Matthew Dicks (DCKMAT004).

October 2, 2021

Contents

1 Estimation **2**
- 1.1 Data . . . 2
- 1.2 Estimation Method . . . 2
- 1.3 Estimation Validation . . . 3
- 1.4 Estimation Results . . . 5

2 Simulation **6**
- 2.1 Simulation Method . . . 7
- 2.2 Simulation Validation . . . 7
- 2.3 Simulation Results . . . 9

3 Code **10**

1 Estimation

This section describes the method that I used to estimate the parameters of the Hawkes Process from the log-likelihood function. It also details the method that I used to test whether the estimated parameters are valid. The final part of this section visualizes the relationship between the event types in the Hawkes Process.

1.1 Data

The data used to estimate the parameters of the Hawkes Process was collected by Jericevich *et al.* (2020) and can be found here. To estimate the parameters I used a day of top-of-book trading data from Naspers (NPN). The day that I used was the 10/07/2019. The reason why I chose to use NPN and the 10/07/2019 was that I wanted the most liquid stock and the day with the most event activity. This would allow me to better fit the parameters of the Hawkes Process. In this dataset, the smallest unit of time was a second. Therefore, trade and quote compacting were done to ensure that each event type can only occur once in a given timestamp. This is a simplifying assumption but it was done to ensure that there are no inter-arrival times of zero which would not be realistic.

After the data was cleaned and compacted, the events were classified into 4 categories. The event types are listed in Table 1. In the top-of-book data, the market orders are the trades that occurred. This is another simplifying assumption because I have assumed that there was never a time where two limit orders matched. The assumption had to be made because the data, in the current form, does not differentiate between these two events. The market orders were classified as buyer-initiated or seller-initiated trades using the Lee/Ready rule (Lee & Ready 1991) and outlined by Theissen (2001).

Event Type	Abbreviation
Limit Order Bid	LOB
Market Order Buy	MOB
Limit Order Ask	LOS
Market Order Sell	MOS

Table 1: Event types used in this analysis.

1.2 Estimation Method

The method that I used to estimate the parameters of the Hawkes Process was the one outlined by Toke & Pomponio (2012). I have used this method as they provide a good definition of the integrated intensity that can be computed numerically. If we define $\mathrm{N(t)} = \{\mathrm{N}^m(\mathrm{t})\}_{m=1}^{M}$ as an M-variate Hawkes Process, where $\mathrm{N}^m(\mathrm{t})$ is the counting process for the event type m. Then by Toke & Pomponio (2012) we have that the log-likelihood of the Hawkes Process can be defined as the sum over the log-likelihoods for each of the m processes. The definition is defined as follows

$$\ln \mathcal{L}\left(\{N(t)\}_{t\leq T}\right) = \sum_{m=1}^{M} \ln \mathcal{L}^m\left(\{N^m(t)\}_{t\leq T}\right) \tag{1}$$

where

$$\ln \mathcal{L}^m\left(\{N^m(t)\}_{t\leq T}\right) = \int_0^T (1-\lambda^m(s)ds + \int_0^T \ln \lambda^m(s)d\mathrm{N}^m(\mathrm{s}) \tag{2}$$

The kernel used was the exponential kernel as defined by Toke & Pomponio (2012). That is the of the following form

$$\phi^{mn}(t-s) = \alpha^{mn} e^{-\beta^{mn}(t-s)} \tag{3}$$

After some simplifications, which can be that can be found in Toke & Pomponio (2012), we get that the log-likelihood for the mth event can be computed using the following formula

$$\ln \mathcal{L}^m \left(\{N^m(t)\}_{t \leq T} \right) = T - \Lambda^m(0,T) + \sum_{l:T_l^m \leq T} \ln \left[\lambda_0^m \left(T_l^m \right) + \sum_{n=1}^{M} \alpha^{mn} R^{mn}(l) \right] \tag{4}$$

where

$$\Lambda^m(0,T) = \int_0^T \lambda_0^m ds + \sum_{n=1}^{M} \sum_{i:T_i^n \leq T} \frac{\alpha^{mn}}{\beta^{mn}} \left(1 - e^{-\beta^{mn}(T - T_i^n)} \right) \tag{5}$$

is the integrated intensity. The recursive function R has the following definition

$$\begin{aligned} R_j^{mn}(l) &= \sum_{T_k^n < T_l^m} e^{-\beta_j^{mn}(T_l^m - T_k^n)} \\ &= \begin{cases} e^{-\beta_j^{mn}\left(T_l^m - T_{l-1}^m\right)} R_j^{mn}(l-1) + \sum\limits_{T_{l-1}^m \leq T_k^n < T_l^m} e^{-\beta_j^{mn}(T_l^m - T_k^n)} & \text{if } m \neq n \\ e^{-\beta_j^{mn}\left(T_l^m - T_{l-1}^m\right)} \left(1 + R_j^{mn}(l-1) \right) & \text{if } m = n \end{cases} \end{aligned} \tag{6}$$

In the 4-variate Hawkes Process there are 36 parameters that need to be estimated. These parameters are as follows $\boldsymbol{\theta} = \left(\lambda_0^1, ..., \lambda_0^4, \alpha^{11}, ..., \alpha^{14}, \alpha^{21}, ..., \alpha^{24}, ..., \alpha^{41}, ..., \alpha^{44}, \beta^{11}, ..., \beta^{14}, \beta^{21}, ..., \beta^{24}, ..., \beta^{41}, ..., \beta^{44} \right)$. To estimate these parameters the maximum-likelihood method used by Toke & Pomponio (2012) was implemented and is as follows

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln \mathcal{L} \left(\{N(t)\}_{t \leq T} \right) \tag{7}$$

To perform this optimization I used the Optim package in Julia. To ensure that all the parameters were all positive real numbers, as per the definition of the Hawkes Process, I used the constrained optimization functionality. I tested a range of optimization methods and I found that LBFGS was the best choice. It was the most efficient optimizer that lead to reasonably good solutions in a relatively short period of time on my computer.

1.3 Estimation Validation

The optimization of the parameters proved to be a very computationally intensive task. Due to the speed and memory resources of my computer I was not able to run the optimization for long enough to ensure that the optimization passed all the convergence tests. However, the parameters that were found were still checked to determine their validity. To perform this validity check I will use Corollary 3.2 from Bowsher (2007), which is as follows

Corollary 3.2: *Let* $\{T_i^{(m)}\}_{i\in\{1,2,...\}}$ *be the sequence of points associated with* $N^m(t)$ *and define* $T_0^{(m)} := 0$, *also* $e_i^m := \int_{T_i^{(m)}}^{T_{i+1}^{(m)}} \lambda^m(s)ds$ *for* $i \in \{0,1,2,...\}$. *Then* $\{e_i\}_{i\in 0,1,2,...}$ *is an i.i.d. sequence of Exponential random variables with mean 1 for m = 1,...,M.*

This corollary provides a nice way to validate the Hawkes Process's parameters by computing the durations and comparing them to the required exponential distribution. Figure 1 compares the distributions of the durations found using the parameters that were obtained by the optimization procedure. Each of the QQ-plots compares the durations of each of the different event types to the Exponential distribution with a mean of 1. As you can see from these plots the durations look to be very close to exponentially distributed with a mean of 1. Even though the optimization procedure was not run for long enough to pass the convergence tests the QQ-plots give evidence to support the fact that the Hawkes Process's parameters are valid estimates.

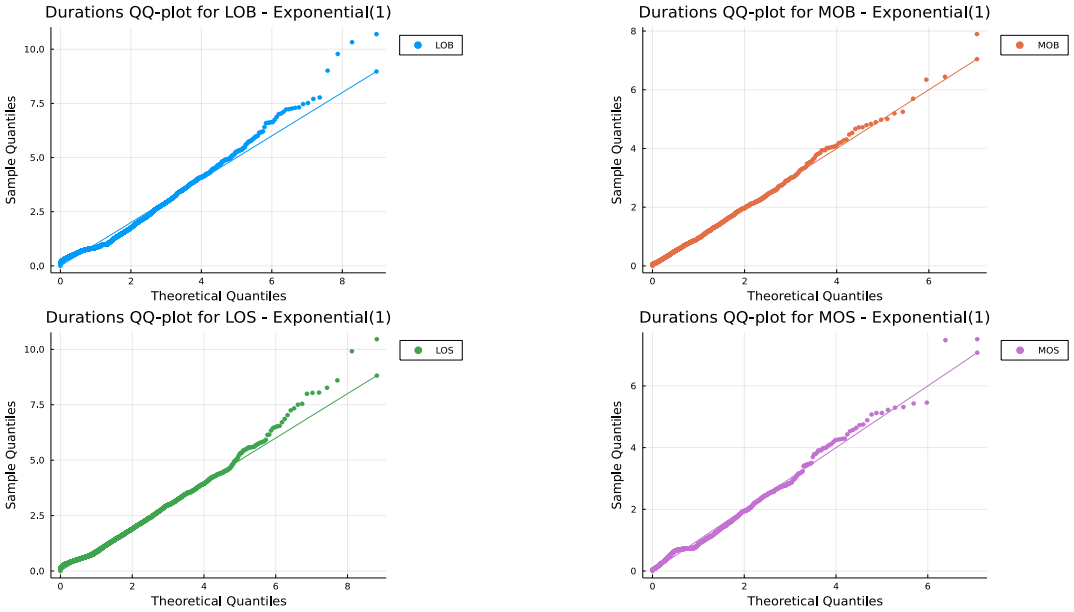

Figure 1: Compares the distribution of the durations of each event type to the Exponential distribution with a mean of 1. Each of the event types durations look to be reasonably Exponentially distributed with a mean of 1.

The corollary not only states that the durations should be exponentially distributed but they should also be i.i.d.. To check how i.i.d. the process is I computed the auto-correlation of the sequence up to a lag of 100. The auto-correlations for each of the 4 event types can be seen in Figure 2. The red bars are the upper and lower bounds of the 95% confidence interval for the significance of the auto-correlation. These values were computed based on the asymptotic assumption that

$$\rho_j \overset{a}{\sim} \mathrm{N}(0, \frac{1}{N}) \tag{8}$$

where ρ_j is the auto-correlation for the j-th lag and N is the number of observations. These plots show that there are some significant correlations in the durations. However, they are not exceedingly large and

do not have a pattern. This indicates that the i.i.d. condition is met reasonably well. This also shows that the parameters could potentially benefit from a longer tuning process where better convergence could lead to better estimates and therefore less significant auto-correlations.

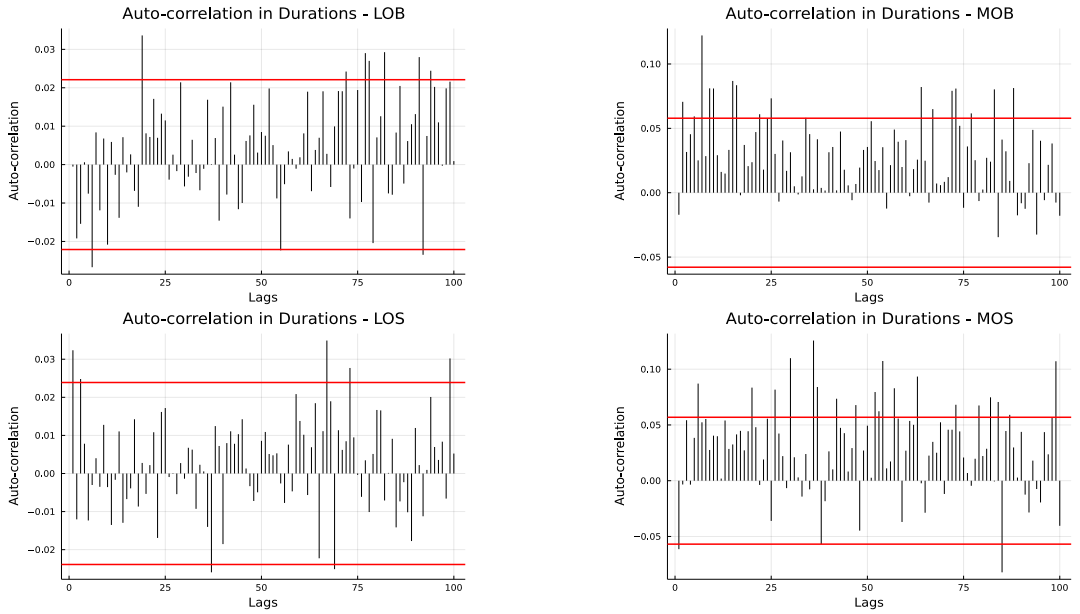

Figure 2: Plots the auto-correlations in the durations to test the i.i.d. condition specified in Corollary 3.2.

The final test that was performed was to check the stationarity of the process. The perform this check I used the test outlined in Toke & Pomponio (2012). This method says that if the spectral radius of the matrix

$$\Gamma = \left(\frac{\alpha^{mn}}{\beta^{mn}} \right)_{m,n=1,\dots,M} \tag{9}$$

is less than 1, then the point process is stationary. This means that the level of market activity for all events considered does not change dramatically during the day (Toke & Pomponio 2012). The spectral radius of the Γ matrix found for the estimated parameters was 0.81. This says that the point process is likely stationary.

1.4 Estimation Results

Section 1.3 gives evidence to suggest that the parameters estimated are valid. This lends credibility to the following visualizations. As Jericevich *et al.* (2021) notes, the branching ratios $\Gamma = \frac{\alpha^{mn}}{\beta^{mn}}$ are the expected number of events of type m caused by a single event of type n, using the exponential kernel described in (3). Jericevich *et al.* (2021) also gives an expression for the half-life of a particular excitation, which gives a sense of how long a particular excitation will last. In this section, I used the same formula they

used for the half-life, which is as follows

$$t_{1/2} = \frac{\log(2)}{\beta^{mn}} \tag{10}$$

Figure 3 plots the expected number of excitations caused by each of the 4 event types as a function of the half-life of the excitations. The size of the circles is proportional to the number of events for a given event type. For all the event types you can see that the largest expected excitations come from self-excitation. For example, the event with the highest number of expected excitations for LOBs are LOBs. This was also found by Large (2007) and Bacry *et al.* (2015). Another artifact to note is that the limit orders do not exhibit any significant cross-excitation effects. However, when considering the market orders there are clear cross-excitations. For example, with MOBs we have that the expected number of excitations of LOBs are approximately 0.27. The MOB also affects the opposite side of the order book where the number of excitations of LOSs caused by a MOB is 0.14. The cross-excitations, on the opposite side of the order book, have relatively longer half-lives when compared with the cross-excitations seen on the same side. A similar effect is also observed for MOSs except that the cross-excitations all have very short half-lives when compared with the cross-excitations for MOBs.

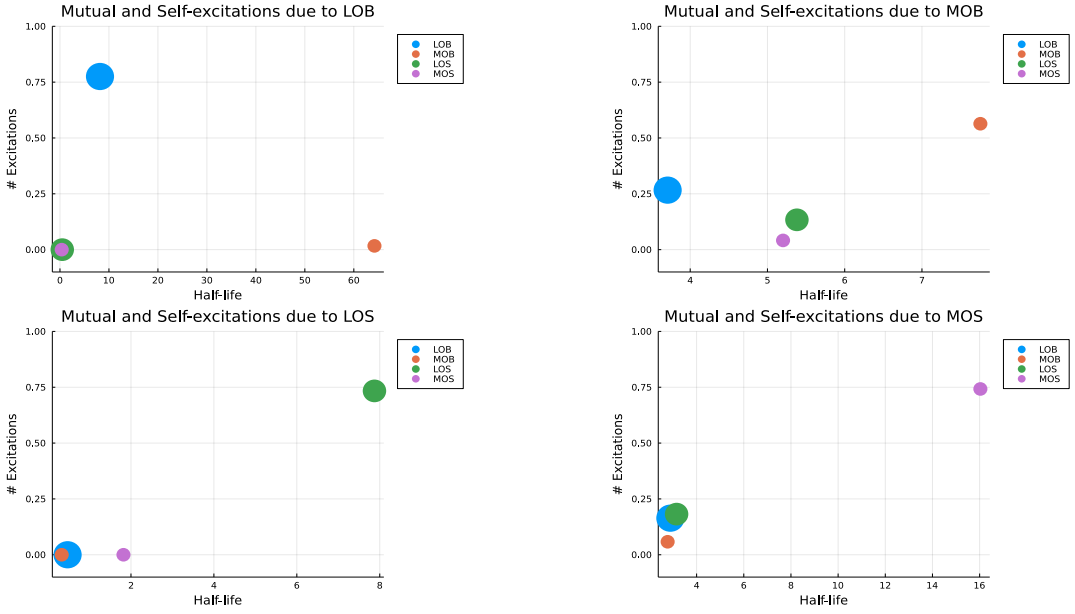

Figure 3: Plots the expected number of excitations caused by each of the 4 event types as a function of the half-life of the excitations. The size of the circles are proportional to the number of events.

2 Simulation

This section describes the method that I used to simulate a 4 dimensional Hawkes Process. It also argues that the event times produced for each event type are valid and thus the simulation was implemented correctly. It also visualizes the simulated Hawkes Process to see how the event time paradigm plays out through the Hawkes Process.

2.1 Simulation Method

To perform the simulation of the Hawkes Process I used the method that was outlined by Toke & Pomponio (2012). Let $\mathcal{U}(0,1)$ be the uniform distributions with upper and lower bounds of 0 and 1, and let [0,T] describe the window where the events will occur. Also let $I^K(t) = \sum_{k=1}^{K} \lambda^k(t)$ be the sum of the first K process' intensities. The following set of steps describes the algorithm used to generate the event times.

1. **Initialization:** Set $i = 0, i^1 = 0, ..., i^M = 0$ and $I^* = I^M(0) = \sum_{m=1}^{M} \lambda_0^m$
2. **First Event:** Generate $U \sim \mathcal{U}(0,1)$ and set $s = -\frac{1}{I^*} \ln U$
   (a) **If** s > T **then** go to step 4
   (b) **Attribution Test:** Generate $D \sim \mathcal{U}(0,1)$ and set $t_1^{n_0} = s$ where n_0 is such that $\frac{I^{n_0-1}(0)}{I^*} < D \leq \frac{I^{n_0}(0)}{I^*}$
   (c) Set $t_1 = t_1^{n_0}$
3. **General Routine:** Set $i^{n_0} = i^{n_0} + 1$ and $i = i + 1$
   (a) **Update maximum intensity:** Set $I^* = I^M(t_{i-1}) + \sum_{n=1}^{M} \alpha^{nn_0}$
   (b) **New Event:** Generate $U \sim \mathcal{U}(0,1)$ and set $s = s - \frac{1}{I^*} \ln U$
   If $D \leq \frac{I^M(s)}{I^*}$
   Then set $t_{i^{n_0}}^{n_0} = s$ where n_0 is such that $\frac{I^{n_0-1}(0)}{I^*} < D \leq \frac{I^{n_0}(0)}{I^*}$, and $t_i = t_{i^{n_0}}^{n_0}$ and go through the general routine again
   Else update $I^* = I^M(s)$ and try a new date at step (b) of the general routine
4. **Output:** Return the simulated process $\{\{t_i^m\}\}_{m=1,...,M}$

To compute the intensity of the mth event I used the following formula detailed in the notes

$$\lambda^m(t) = \lambda_0 + \sum_{n=1}^{M} \int_{-\infty}^{t} \alpha^{mn} e^{-\beta^{mn}(t-s)} dN^n(s) \tag{11}$$

$$= \lambda_0 + \sum_{n=1}^{M} \sum_{t_k^n < t} \alpha^{mn} e^{-\beta^{mn}(t-t_k^n)} \tag{12}$$

The parameters that I used to run the simulation were the same as those estimated during the estimation process. This will allow for a comparison of the event counts between the simulated process and the observed process. It will also give an indication of the validity of the parameters because if the counts are similar this would give some evidence that the estimated process describes the observed event times reasonably well for the given day.

2.2 Simulation Validation

As I did to validate the parameters of the estimation I will once again use Corollary 3.2 from Bowsher (2007) to validate the event times generated from the simulation. Figure 4 compares the distribution of

the durations to the Exponential distribution with a mean of 1. It can be seen from the plots that the durations generated by the simulation are very close to the required distribution. Given how hard it is to check if a simulation has run correctly this gives some evidence that the implementation is correct. I also checked the i.i.d. assumption in the same way as I did in the estimation section. Figure 5 shows the auto-correlations of the durations and plots the required 95% confidence interval. From this figure, it can be seen that the durations do not have very significant auto-correlations (except for a few lags) providing further evidence that the simulation was implemented correctly. Given that the parameters used were the ones estimated from the trading day and have shown to have a spectral radius of less than 1, we can also say that the process that has been simulated is stationary.

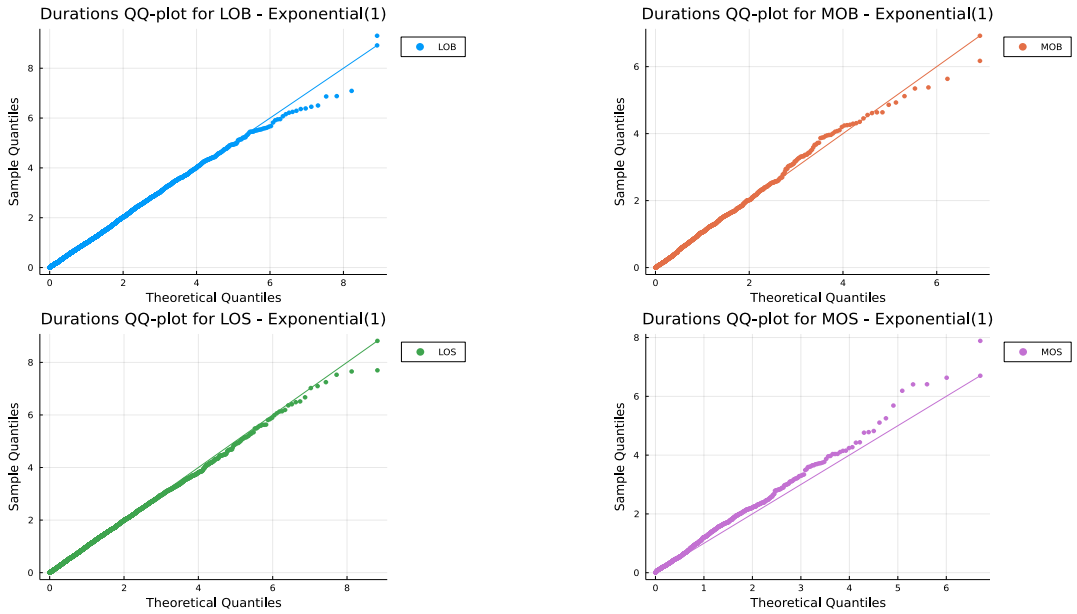

Figure 4: Compares the distribution of the simulated durations to the Exponential distribution with a mean of one.

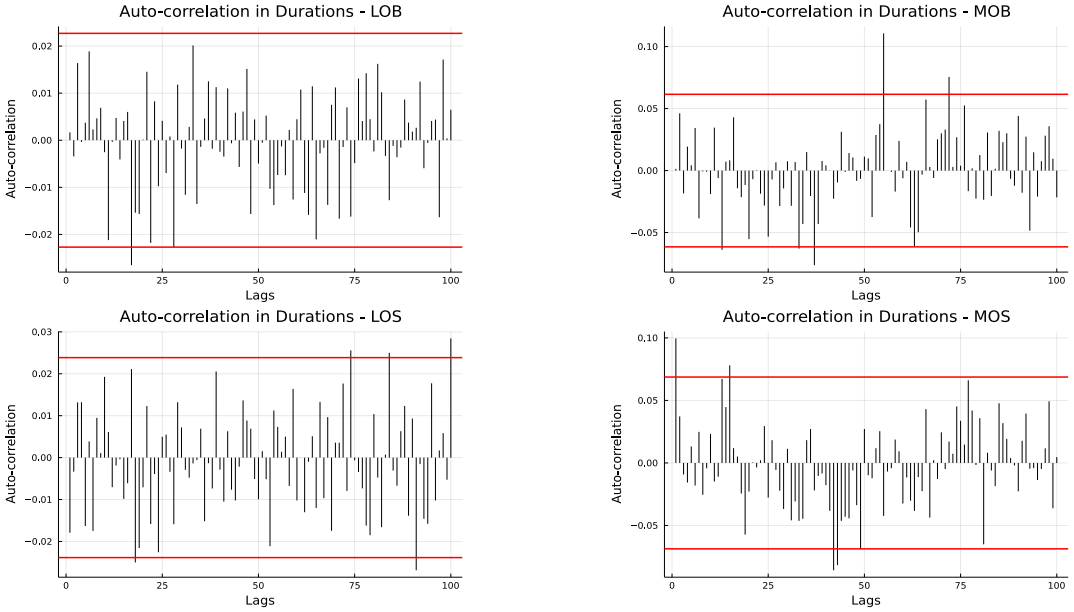

Figure 5: Plots the auto-correlations for durations generated by the simulation. The red bar represents the 95% confidence interval.

One advantage of using the parameters generated from the estimation process is that I can compare the actual counts of the events observed on the JSE and the counts realized from the simulation. This will provide an additional way to validate the simulation and the estimation process. Table 2 compares the observed counts to the simulated counts. The simulation's counts look to have reasonably recovered the counts of the observed data. However, for LOBs and the MOSs we have that the simulation produced 390 and 373 fewer events for these event types respectively. This shows that even though the simulation has produced valid results there looks to be some areas where it could improve. The difference could potentially be due to a bug in the implementation or it could be due to the estimates not being optimal in the sense that the convergence of the optimization procedure did not pass all the tests and still exhibited a large gradient. Increasing the optimization time could produce even closer counts. Despite the differences, the counts generated from the simulation look to have recovered the different behaviors of the event types. For example, the limit orders have far larger counts when compared with the market orders.

Event Type	Observed Counts	Simulated Counts
Limit Order Bid	7865	7457
Market Order Buy	1148	1016
Limit Order Ask	6733	6754
Market Order Sell	1188	815

Table 2: The number of events for each event type for the observed data and the simulated data.

2.3 Simulation Results

Given that the simulation has shown to have produced reasonably valid results ensures that the counting process' events are worth investigating. I have visualized the counting process and the events of each of

the counting processes in Figure 6. The plot on the left of the figure shows approximately the first 500 seconds of the counting processes. In the figure, the small circles represent when the events occurred. As you can see initially the market is dominated by lots of quotes before there are any market orders. The plot on the right of Figure 6 shows a little more clearly when each of the events occurred without the counts dominating the visualization. In this plot, you can more clearly see the event time paradigm come into play as no calendar time or loop structure is generating these events. The time between events is exponential and in these plots, you can see how this behavior arises. This is a clear example of how, at the microstructure level, events are generated by a self and mutually-exciting process. The self-excitation effect can be seen by the LOS event type on the right-hand plot. From the start to the end of this period there is an acceleration in the number of LOS events. It is also worthwhile to point out that each of the events start at different times, which indicates that these events are being triggered by the occurrence of other events.

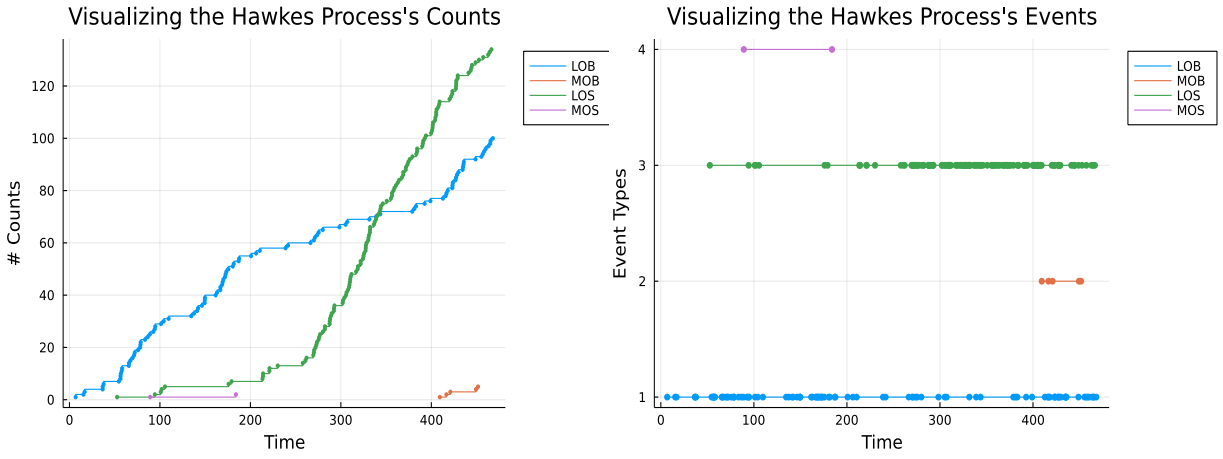

Figure 6

3 Code

The GitHub repository for this lab exam can be found here.

References

1. Bacry, E., Mastromatteo, I. & Muzy, J.-F. Hawkes processes in finance. *Market Microstructure and Liquidity* **1,** 1550005 (2015).
2. Bowsher, C. G. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics* **141,** 876–912 (2007).
3. Jericevich, I., Chang, P. & Gebbie, T. Comparing the market microstructure between two South African exchanges. *arXiv preprint arXiv:2011.04367* (2020).
4. Jericevich, I., Chang, P. & Gebbie, T. Simulation and estimation of a point-process market-model with a matching engine. *arXiv preprint arXiv:2105.02211* (2021).
5. Large, J. Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets* **10,** 1–25 (2007).
6. Lee, C. M. & Ready, M. J. Inferring trade direction from intraday data. *The Journal of Finance* **46,** 733–746 (1991).
7. Theissen, E. A test of the accuracy of the Lee/Ready trade classification algorithm. *Journal of International Financial Markets, Institutions and Money* **11,** 147–165 (2001).
8. Toke, I. M. & Pomponio, F. Modelling trades-through in a limit order book using Hawkes processes. *Economics* **6** (2012).