Censored EM algorithm for Weibull mixtures: application to arrival times of market orders

Markus Kreer, Ayşe Kızılersıı, and Anthony W. Thomas

1CSSM, Faculty of Sciences, School of Physical Sciences, Department of Physics, University of Adelaide, 5005, Adelaide, Australia.
2CAMPUSERVICE GmbH, Servicegesellschaft der Johann Wolfgang Goethe-Universität Frankfurt, Rossertstrasse 2, 60323 Frankfurt am Main, Germany
3Feldbergshule, Oberhochstädter Str., 20, 61440, Oberrursel (Taunus), Germany

In a previous analysis the problem of “zero-inflated” time data (caused by high frequency trading in the electronic order book) was handled by left-truncating the inter-arrival times. We demonstrated, using rigorous statistical methods, that the Weibull distribution describes the corresponding stochastic dynamics for all inter-arrival time differences except in the region near zero. However, since the truncated Weibull distribution was not able to describe the huge “zero-inflated” probability mass in the neighbourhood of zero (making up approximately 50% of the data for limit orders), it became clear that the entire probability distribution is a mixture distribution of which the Weibull distribution is a significant part. Here we use a censored EM algorithm to analyse data for the difference of the arrival times of market orders, which usually have a much lower percentage of zero inflation, for four selected stocks trading on the London Stock Exchange.

I. INTRODUCTION

Electronic Order Book (EOB) trading is now common to most stock exchanges. A set of EOB data from the FTSE for the period June to September 2010 was described and partly analysed in Ref.[1]. In this data set timestamps of changes in the EOB were given in milliseconds and because of the huge volume of EOB trading by (ultra-)high frequency trading algorithms, taking place on a microsecond scale, differences of timestamps appeared to be “zero-inflated”. Because of this rounding error, all differences in timestamps, \( \Delta t \), are mapped to our actual data set as follows

\[
\Delta t \in (0.0, 0.5) = I_1 \rightarrow 0 = c_1 \\
\Delta t \in (0.5, 1.5) = I_2 \rightarrow 1 = c_2 \\
\Delta t \in (1.5, 2.5) = I_3 \rightarrow 2 = c_3 \\
\text{etc.}
\]

where the integer numbers \( c_1, c_2, \ldots \) are the rounded time stamp differences to the precision of milliseconds.

In Ref.[1] it was demonstrated that sets of non-negative time differences between subsequent market orders (MO) with \( \Delta t > 10 \) milliseconds describe a random variable \( X \) that could be fitted to parametric distributions, such as a left-truncated Weibull distribution, which passed a rigorous set of goodness-of-fit tests, such as Kolmogorof-Smirnov, Anderson-Darling, Cramer von Misses and Kuiper tests[2].

Here we analyse the entire set of observations of the random variable \( X \), including the “zero-inflated” ones, by starting from a Weibull distribution for the whole range

\[
f(x|\alpha_i, \beta_i) = \frac{\beta_i}{\alpha_i^\beta_i} \left( \frac{x}{\alpha_i} \right)^{\beta_i-1} e^{-\left( \frac{x}{\alpha_i} \right)^{\beta_i}}, \quad (1)
\]

where \( \alpha_i > 0 \) is the scale parameter and \( \beta_i > 0 \) the shape parameter. Note that for \( \beta_i = 1 \) we recover the exponential distribution. We denote the parameter vector by \( \theta_i = (\alpha_i, \beta_i) \) and use the subscript index \( i \) to denote various Weibull or exponential distributions with different parameter vectors \( \theta_i \).

Now, the attempt to analyse the entire data set including the “zero-inflated” part (which sometimes was even more than half of the observed data set) requires a different procedure. One promising approach is to apply interval censoring to the observed data for the intervals of small time differences \( \Delta t \): we use the above intervals \( I_1, I_2, I_3, \ldots \) together with the information of how many observations belong to them, i.e. \( N_1, N_2, N_3, \ldots \). In this notation we can write an observed sample of a positive random variable \( X \) as

\[
x_1, x_2, \cdots, x_n, x_{n+1}, x_{n+2}, \ldots, x_N = x_1, x_2, \ldots, x_n, c_1, c_1, \ldots, c_L, c_L, \ldots, c_L, (2)
\]

where, after grouping, all observations with index bigger than \( n \) have been censored (with \( L \) censoring intervals).

Despite censoring, the random variable \( X \) itself will be assumed to come from a mixed distribution with a mixture of \( M \) components. As a consequence of the findings in Ref.[1], we expect that a significant component of this mixed distribution should come from a Weibull distribution. Similar results where found in Ref.[3]: the authors found that the differences between subsequent MO’s for 30 DJIA stocks from the NYSE in October 1999 can be well described by Weibull distributions. Thus we want to investigate here mixtures of the following kind

\[
f(x|\theta) = \sum_{i=1}^{M} \Pi_i f_i(x|\theta_i), \quad (3)
\]

where the weights, \( \Pi_i \), satisfy \( \Pi_i \geq 0 \) with \( \sum_{i=1}^{M} \Pi_i = 1 \) and the probability density functions \( f_i(\cdot|\theta_i) \) are given by Eq.(1), where the parameters \( \theta_i \) need to be estimated.
II. CENSORED EM ALGORITHM FOR MIXTURES IN A NUTSHELL

The EM algorithm is an iterative procedure where an expectation-step (E-step) tries to estimate the unobservable indicators $z_{ij}$ and $\tilde{z}_{i\ell}$, respectively, and then uses the result in the maximization step (M-step) to estimate the parameters of the mixtures by an MLE. After the M-step there is another E-step and so on. Here we follow closely Ref.[9], where a hint is given that this iteration converges for certain function families (including exponential and Weibull distribution functions) in a certain limited parameter range. In the following sections the integer index $k$ denotes the number of the iteration.

A. The E-step

Using the above notation and assuming that the parameter vector $\theta_i^{(k-1)}$ for each mixture component $i$ are known from a previous step, Eq.(3.3) in the Ref.[9] provides the following term for the uncensored observations, $j = 1, ..., n$ for the mixture component $i$

$$z_{ij}^{(k)} = \frac{f_i(x_j|\theta_i^{(k-1)})}{\sum_{\ell=1}^{M} f_i(x_j|\theta_i^{(k-1)})} \cdot \Pi_i^{(k-1)}$$

and for the censored observations in the censoring intervals $I_{\ell} = (\xi_{\ell-1}, \xi_{\ell})$ with $\ell = 1, 2, ..., L$. Eq.(3.4) in the Ref.[9] provides the following expression for the mixture $i$th component

$$z_{i\ell}^{(k)} = \frac{\int_{I_{\ell}} dy f_i(y|\theta_i^{(k-1)})}{\sum_{\ell=1}^{M} \int_{I_{\ell}} dy f_i(y|\theta_i^{(k-1)})} \cdot \Pi_i^{(k-1)}$$

B. The M-step

The sample size, $N = n + \sum_{\ell=1}^{L} N_{\ell}$, is the number of all observations (uncensored and censored) and the new weights after the E-step are given by the following update-rule (Eq.(3.7) of Ref.[9]):

$$
\Pi_i^{(k)} = \frac{1}{N} \sum_{j=1}^{n} \frac{f_i(x_j|\theta_i^{(k-1)})}{\sum_{i=1}^{M} \Pi_i^{(k-1)} f_i(x_j|\theta_i^{(k-1)})} \cdot \Pi_i^{(k-1)}
+ \frac{1}{N} \sum_{\ell=1}^{L} N_{\ell} \int_{I_{\ell}} dy f_i(y|\theta_i^{(k-1)}) \sum_{i=1}^{M} \frac{f_i(x_j|\theta_i^{(k-1)})}{\Pi_i^{(k-1)} f_i(x_j|\theta_i^{(k-1)})} \cdot \Pi_i^{(k-1)}.
$$

Note that this update-rule is just same as the following formula

$$\Pi_i^{(k)} = \frac{1}{N} \left( \sum_{j=1}^{n} z_{ij}^{(k)} + \sum_{\ell=1}^{L} N_{\ell} z_{i\ell}^{(k)} \right)$$

Now, the following expression needs to be maximised with respect to the parameter vectors $\theta_i^{(k)}$, where the index $i$ refers to the mixture and $k$ to the actual number in the iterative procedure
Here, the conditional density function on the censoring interval \( \mathcal{I}_\ell \) is given by Ref.\cite{9} Eq.\( \text{(3.5)} \)

\[
h_i(y|c_\ell, \theta_i^{(k-1)}) = \frac{f_i(y|\theta_i^{(k-1)})}{\int_{\mathcal{I}_\ell} dy \ f_i(y|\theta_i^{(k-1)})}
\]  

(9)

The MLE equations will be obtained by differentiating Eq.\( \text{(8)} \) with respect to the components of the parameter vector \( \theta_i^{(k)} \). Note that in this expression terms like \( \sum_{i,j} z_{ij} \log \Pi_{ik} \) and the normalization condition \( \mu \cdot \left( \sum_i \Pi_{ik} - 1 \right) \) are not given because they depend only on the parameter vector \( \theta_i^{(k-1)} \) from the previous iteration and are irrelevant for the maximization with respect to the parameter vector \( \theta_i^{(k)} \). Note also that the maximization problem decouples into independent maximization problems for each individual probability distribution.

III. IMPLEMENTATION OF THE CENSORED WEIBULL MIXTURES

For simplicity we now consider a 2-component mixture consisting of an exponential distribution, denoted by \( i = 1 \), and a general Weibull distribution, denoted by \( i = 2 \), as given in Eq.\( \text{(1)} \). Note that here the \( \alpha \) and \( \beta \)

\[
0 = \sum_{j=1}^{n} z_{ij} \left[ -1 + \frac{x_j}{\alpha_2^{(k)}} \right] \beta_2^{(k)} + \sum_{\ell=1}^{L} N_\ell z_{2\ell} \left\{ -1 + \frac{\alpha_2^{(k-1)}}{\alpha_2^{(k)}} \right\} \beta_2^{(k)} \Gamma \left( \frac{\beta_2^{(k)}}{\beta_2^{(k-1)} + 1, \zeta_{k-1}} \right) \frac{1}{e^{-\zeta_{k-1}} - e^{-\zeta_k}}
\]

(11)

and likewise for \( \frac{\partial}{\partial \beta_2^{(k)}} \)

\[
0 = \sum_{j=1}^{n} z_{ij} \left[ \frac{1}{\beta_2^{(k)}} + \log \frac{x_j}{\alpha_2^{(k)}} - \frac{x_j}{\alpha_2^{(k)}} \log \frac{x_j}{\alpha_2^{(k)}} \right]
\]

\[
+ \sum_{\ell=1}^{L} N_\ell z_{2\ell} \left\{ \frac{1}{\beta_2^{(k)}} + \log \frac{\alpha_2^{(k-1)}}{\alpha_2^{(k)}} + \frac{D_{2\ell}^{(k)}}{e^{-\zeta_{k-1}} - e^{-\zeta_k}} \right\}
\]

(12)

have a subscript \( i \) for the mixture and a superscript \( (k) \) for the iteration step in the EM algorithm. The expression corresponding to Eq.\( \text{(8)} \) is given in Appendix A. Our algorithmic results extend the results as given by Ref.\cite{10} for exponential mixtures. It is clear how to modify our computations for arbitrary mixtures, e.g. \( (p + r) \) mixtures consisting of \( p \) exponentials and \( r \) Weibulls, with \( p, r = 0, 1, 2, ... \)

A. MLE equations for M-step

Maximising the expression in Appendix A, Eq.\( \text{(A1)} \), with respect to the first parameter, \( \alpha_1^{(k)} \) leads for the exponential distribution (index \( i = 1 \)) to an explicit solution

\[
\alpha_1^{(k)} = \frac{\sum_{j=1}^{n} z_{ij}^{(k)} x_j + \sum_{\ell=1}^{L} N_\ell z_{1\ell}^{(k)} C_{1\ell}^{(k-1)}}{\sum_{j=1}^{n} z_{ij}^{(k)} + \sum_{\ell=1}^{L} N_\ell z_{1\ell}^{(k)}}
\]

(10)

where the quantity \( C_{1\ell}^{(k-1)} \) is defined in Appendix A, Eq.\( \text{(A2)} \).

The MLE equations for the Weibull distribution (index \( i = 2 \)) are obtained by computing \( \frac{\partial}{\partial \alpha_2^{(k)}} \) and equating the expression to 0,

The auxiliary functions \( D_{2\ell}^{(k)} \), as well as \( \zeta_\ell \), are defined in the Appendix A. Reference \cite{11} has obtained similar results for the complete finite mixture of Weibull distributions albeit without the censoring terms displayed here.
B. Algorithmical implementation

We use one censoring interval only, $I_1 = (0, 0.5)$ and $L = 1$ to handle our zero-inflated data sets (see Ref.[1]). Also in the spirit of Ref.[12], as a practical approximation to simplify the solution of these non-linear equations, we use a self-consistency assumption by setting the ratios $\alpha_2^{(k-1)}/\alpha_2^{(k)}$ and $\beta_2^{(k-1)}/\beta_2^{(k)}$ equal to 1 in the MLE equations for the Weibull parameters. This considerably simplifies the MLE equations, Eqs. (11)–(12), and in our experience leads to rapid convergence in most cases.[13] For the EM-algorithm we proceed as follows[14]:

1. At $k = 0$ initialize mixture weights $\Pi_i^{(0)}$ and initial values $\alpha_1^{(0)}, \alpha_2^{(0)}, \beta_2^{(0)}$

2. Compute $z_{ij}^{(k)}, \tilde{z}_{i\ell}^{(k)}, \Pi_i^{(k)}$ using Eq。(5), (6), (7) for $k = 1, 2, ...$

3. Compute $\alpha_1^{(k)}$ using Eq。(10) for exponential distribution for $k = 1, 2, ...

4. Compute $\alpha_2^{(k)}$ using Eq。(11) for Weibull distribution putting here only $\beta_2^{(k)} = \beta_2^{(k-1)}$ for $k = 1, 2, ...

5. Compute $\beta_2^{(k)}$ using Eq。(12) for Weibull distribution for $k = 1, 2, ...

6. Compute the current log-likelihood from Eq。(3)–(4)

7. If the absolute value of “log-likelihood at step $k$ minus log-likelihood at step $k-1$” is bigger than $\varepsilon > 0$, then put $k \to k + 1$ and go back to step 2., otherwise terminate.

In our experience this version of the censored EM algorithm (based on the MLE equations) converges sufficiently fast to a desired maximum solution for suitable initial conditions. We have taken $\varepsilon = 10^{-5}$ and start with equal mixtures setting $\beta = 1$ and take values of $\alpha$ motivated by our previous study Ref.[1]. When testing the algorithm, its results have been cross-checked by the corresponding algorithm which uses a direct maximisation of the objective function given in Eq。(8) in the M-step. Usually the results obtained both way agree fairly well. In Ref.[9] it is claimed that there is local convergence almost surely.

IV. ANALYSIS OF MO ARRIVAL TIMES AND MODEL SELECTION

A. A first approach: “naive” analysis of entire data set

Here we take all time stamps of a given stock from 1st June to 30th September 2010 into one large sample and fit various mixture models to the arrival times (i.e. the difference of subsequent timestamps) using the algorithm described above. This analysis would be sensible if the stochastic process were stationary and the sample data were independent identically distributed (i.i.d). In Table I we provide the log-likelihood per data point to obtain a first idea about the best model. The percentage numbers in round brackets denote the proportion of those data points which have been censored. Note that the quantity “log-likelihood per data point”, or “average log-likelihood”, corresponds to a negative Shannon entropy per event because for large sample size $N$ and i.i.d. data we have under the usual consistency property of the maximum likelihood estimator

$$\frac{1}{N} \log L \sim -\sum \log p = -\int \log p \, dp$$

The physical meaning of this quantity is the “information content” or “surprisal” when a new MO enters the EOB.

| Model | RIO (2.5%) | BARC (2.3%) | RRLN (2.5%) | ABFLN (2.1%) |
|-------|------------|-------------|-------------|--------------|
| 1 exp + 1 wbl | 4 | -8.357 | -8.294 | -9.510 | -10.074 |
| 0 exp + 2 wbl | 5 | -8.349 | -8.288 | -9.492 | -10.060 |
| 3 exp + 0 wbl | 5 | -8.422 | -8.394 | -9.792 | -10.323 |
| 2 exp + 1 wbl | 6 | -8.350 | -8.290 | -9.494 | -10.061 |
| 1 exp + 2 wbl | 7 | -8.348 | -8.288 | -9.489 | -10.057 |
| 4 exp + 0 wbl | 7 | -8.360 | -8.300 | -9.511 | -10.081 |
| 0 exp + 3 wbl | 8 | -8.348 | -8.288 | -9.489 | -10.054 |
| 3 exp + 1 wbl | 8 | -8.349 | -8.289 | -9.491 | -10.055 |
| 5 exp + 0 wbl | 9 | -8.351 | -8.291 | -9.497 | -10.063 |

We see from Table I that all models seem to yield very similar Shannon entropy measures for an individual stock. Also, depending on the trading activity, the entropies differ: for a heavily traded stock such as RIO or BARC the “surprise” is lower than for less actively stocks such as RRLN or ABFLN. A model selection criterion here would be the model with least entropy and thus a mixture of 3 Weibull distributions seems to be the best choice.

However, we know from Ref.[1] that the data are only i.i.d. for smaller subsamples and thus the scale parameters in the exponential and Weibull distribution will exhibit a time dependence. When looking at smaller samples it became customary for model selection rather looking at the above Shannon entropy to put the sample size in relation to the degrees of freedom (dof) (see e.g. Ref.[15]). We have decided to use the Bayesian Information Criterion (BIC) (Ref.[16], Ref.[15])

$$\text{BIC} = -2 \log L + d \log N$$

where $L$ are likelihoods and $d$ is the number of degrees of freedom of the individual model. From Table I we can generate hypothetically Table II by using Eq。(13) with
\( N = 200 \) for “larger stocks”, with higher trading activity (RIO, BARC), and \( N = 100 \) for “smaller stocks”, with lower trading activity (RRLN, ABFLN). These values for \( N \) correspond to time intervals of approximately 10 minutes and we have seen in Ref.[1] that these time intervals yield samples for which the assumption of stationarity of the data set can be somehow justified. We clearly see that now mixtures with low dof are favored, in particular the mixture “1 exponential + 1 Weibull”. Note also that the suggestion of Ref.[17] to model the arrival time distribution as a suitable mixture of exponential waiting times, will be excluded in the model selection using BIC by a too high value for the dof.

### Table II: Expected BIC from Table I with different sample size \( N \)

| Model          | dof | RIO  (2.5%) | BARC (2.3%) | RRLN (2.5%) | ABFLN (2.1%) |
|----------------|-----|------------|------------|------------|-------------|
| 1 exp + 1 wbl  | 4   | 3383.99    | 3388.79    | 1923.19    | 2035.99     |
| 0 exp + 2 wbl  | 5   | 3366.09    | 3341.69    | 1924.89    | 2038.49     |
| 3 exp + 0 wbl  | 5   | 3395.29    | 3384.09    | 1984.89    | 2091.09     |
| 2 exp + 1 wbl  | 6   | 3371.79    | 3347.79    | 1930.59    | 2043.99     |
| 1 exp + 2 wbl  | 7   | 3376.29    | 3352.29    | 1934.89    | 2048.49     |
| 4 exp + 0 wbl  | 7   | 3381.99    | 3357.99    | 1940.59    | 2053.39     |
| 0 exp + 3 wbl  | 8   | 3381.59    | 3357.59    | 1940.19    | 2053.19     |
| 3 exp + 1 wbl  | 8   | 3381.99    | 3357.99    | 1940.59    | 2053.39     |
| 5 exp + 0 wbl  | 9   | 3388.08    | 3364.08    | 1947.08    | 2060.28     |

### B. A second approach: Analysis of stationary subsamples

The previous analysis was naive as we assumed the entire data sample would consist of i.i.d. random variables. This is clearly not the case. As already noticed in Ref.[1] the volume of trading changes a great deal during a trading day, so that the scale parameter \( \alpha \) must also vary. However, in Ref.[1] Kızılersi et al. argued that for “small” subsamples the assumptions of stationarity and i.i.d. might be expected to be justified. We take as subsample size \( N = 200 \) for “larger stocks” with higher trading activity (RIO, BARC) and \( N = 100 \) for “smaller stocks” with lower trading activity (RRLN, ABFLN). Motivated by Table II, our candidates for possible models are the following mixtures

- 1 exp + 1 wbl (dof=4)
- 0 exp + 2 wbl (dof=5)
- 3 exp + 0 wbl (dof=5)
- 2 exp + 1 wbl (dof=6)

To quote Ref.[18] “[t]he practice of using the same data set to select a best-fitting model and to assess the significance of model parameter estimates or interpret the model structure is based on the often implicit assumption that the selected model is the true model that generated the data [...]. However, this assumption does not hold in general. The sampling error related to model selection is ignored if the same data are used for inference.” Thus, we separate the task of model selection from the best fit of the parameters.

#### 1. Model selection and bootstrapping

For the model selection we take for every trading day in the months of June and July 2010 a random time stamp for each individual stock. From this time stamp onward we take 200 successive time stamps for the big stocks (RIO and BARC) and 100 successive time stamps for the small stocks (RRLN and ABFL). For each of these original samples we generate additional 999 bootstrap samples out of the original sample (e.g. for more details on bootstrapping the standard reference [19]). For each ensemble of 1000 bootstrap samples we run the censored EM-algorithm and compute the log-likelihood and the BIC. Hence, for each model we have obtained a BIC distribution which is approximately normal. We then perform a Welch t-test with 5% confidence level on the following hypothesis

“Can the alternative model beat 1 exp + 1 Weibull using BIC?”

We then count the success rate for the winning distribution. Our results are depicted in Table III, displaying the proportion of winnings. Our first intuition is confirmed: the mixture “1 exponential + 1 Weibull” is the clear winner.

### Table III: BIC-winners from bootstrapping ensembles in June and July 2010

| Model          | dof | RIO | BARC | RRLN | ABFLN |
|----------------|-----|-----|------|------|-------|
| 1 exp + 1 wbl  | 4   | 0.77| 0.77 | 0.77 | 0.76  |
| 0 exp + 2 wbl  | 5   | 0.09| 0.14 | 0.09 | 0.12  |
| 3 exp + 0 wbl  | 5   | 0.14| 0.09 | 0.14 | 0.12  |
| 2 exp + 1 wbl  | 6   | 0   | 0    | 0    | 0     |

#### 2. Results for the preferred model: “1 exponential + 1 Weibull”

In this subsection we use the convention that for the exponential contribution of the mixture \( \beta_1 = 1 \) will be suppressed and for the Weibull contribution we write \( \beta \) rather than \( \beta_2 \) for easier reading. In Table IV the results of the estimated parameters are summarised. We find for the “complete” data samples that the Weibull shape parameter \( \beta \) takes on a universal value of approximately 0.57 as already found in Ref.[1] (where using a left-truncation was the way of handling the “zero-inflated” data).
TABLE IV: Weibull component for “1 exp + 1 Weibull”

|          | N \(\log L \pm \Delta \log L\) | weight | median \(\beta\) | \(\beta \pm \Delta \beta\) | No. samples |
|----------|----------------------------------|--------|------------------|-----------------------------|-------------|
| RIO      | -8.29±0.63                       | 0.82   | 0.55             | 0.57 ± 0.11                 | 3107        |
| BARC     | -8.24±0.72                       | 0.81   | 0.57             | 0.58 ± 0.11                 | 3346        |
| RRLN     | -9.35±0.83                       | 0.78   | 0.48             | 0.50 ± 0.12                 | 535         |
| ABFLN    | -9.91±0.78                       | 0.81   | 0.47             | 0.49 ± 0.12                 | 291         |

From Table IV we see that the median of the Weibull scale parameter \(\alpha_2\) varies for different stocks (depending on their trading activity), whereas the median of the exponential scale parameter \(\alpha_1\) seems to be the same for different stocks. From Table V we suspect a stronger time dependence for the Weibull scale parameter. To investigate this time dependence, we divide the trading hours from 9:00 to 17:30 UK summer time in intervals of 10 (respectively 30) minutes for RIO and BARC (RRLN and ABFLN respectively) and average the values of \(\alpha_2\), \(\alpha_1\) and \(\beta\) over the trading days from June to September 2010. Due to this partitioning of the data, the sample size will vary significantly, depending on the time of the day, with \(N \gg 200\) at some times and \(N \ll 200\) at some other times. As we have already remarked in Ref.[1] the Weibull scale parameter \(\alpha_2\) will exhibit a strong time dependence during the trading day. The more actively a stock is traded, the smaller the Weibull scale parameter \(\alpha_2\) will be. We see from Figure 1 that the typical scale parameter for the big stocks RIO and BARC is well below 10 seconds, and both curves as function of time are nearly identical. This can be explained by index arbitrage in the FTSE100, which requires trading in big stocks such as RIO and BARC at the same time to exploit the arbitrage. Obviously at lunch time there is less activity and the \(\alpha_2\) becomes larger.

Since the Fisher information matrix is not diagonal for the MLE problem for Weibull distributions a bias in the estimated scale parameter \(\alpha_2\) will also result in a bias of the estimated shape parameter \(\beta\). Thus, we see in Figure 2 that the value of \(\beta\) becomes slightly larger at lunch time, when the estimated value of \(\alpha_2\) is larger than that found at the opening or closing of trading hours. Despite this, the high level of stability found for \(\beta\) suggests that it is reasonable to assume that the true shape parameter is universal with \(\beta = 0.57\) as conjectured in Ref.[1].

Finally we point the reader’s attention to Fig. 3 which displaying the time dependence of the scale parameters \(\alpha_1\) of the exponential contribution and Fig. 4 displaying the time dependence of the mixture weights: the time
dependence of \(\alpha_1\) is mainly in the region of 10 to 20 milliseconds for all stocks with exceptions of the “smallest” and sometimes illiquid stock ABFLN. Also the weights of this exponential contribution in the two-component mixture is nearly independent of time and consistently around 20 %, valid for all stocks.

V. CONCLUSION

The censored EM-algorithm in combination with a bootstrapping argument applied to the Baysian Information Criterion (BIC) allows us to choose as a model for the MO arrival times a two-component mixture distribution consisting of “1 exponential + 1 Weibull”. Of course, this conclusion is not applicable in the exceptional case of extraordinary trading activity in a stock...
when critical information is disclosed to the market participants. The first component of this mixed distribution is an exponential distribution with a relative weight of approximately 20% and a rather short scale parameter, in the range of 10 to 20 milliseconds. This result is independent of the stock under consideration and almost constant during the trading day. The second component, with a relative weight of approximately 80%, is a Weibull distribution for which the scale parameter varies intra-day with trading activity and lies typically between 1000 and 25000 milliseconds, albeit with universal shape parameter $\beta \approx 0.57$.

This result can be understood with a view to a typical stock exchange computer architecture (e.g. Ref.[20]): market orders are captured via various order gateways which forward the electronic orders to the so called accumulator, where timestamps are given. Low latency order gateways typically cater for high-frequency traders (e.g. algorithmic hedge funds) whose buy- and sell-orders are generated by computers located in the stock exchange’s immediate neighbourhood to minimize the transmission time. Here an exponential waiting time between these signals (with a short time scale parameter) is to be expected resulting in a Poisson process for high-frequency buy- and sell-orders. On the other hand, VSAT gateways cater for stock brokers whose customers include both institutional and private clients and whose orders are usually transmitted via an internet connection. It has been shown that internet traffic sends information as TCP “parcels” with arrival times that can be described by Weibull distributions with a shape parameter less than 1 (see Ref.[21], Ref.[22]). In Ref.[21] it was shown that a shape parameter $\beta = 0.569$ provided an excellent fit.
of how to fit a waiting time distribution consisting of a finite number of exponential waiting time distributions to the observed tick-by-tick data. Although exponential mixtures with more than 5 components seem to describe our actual data sets very well, they are excluded by the BIC due to their too high dof-value.

Appendix A: Expression of objective function in M-step and further auxiliary functions

First define the censoring intervals $I_\ell = [\xi_{\ell-1}, \xi_\ell)$. For simplicity introduce a transformed quantity $\zeta_\ell = (\xi_\ell / \alpha_1^{(k-1)})^{\beta_1^{(k)}}$. Then we find after some lengthy computations in the spirit of section 2 the following version of Eq.(8)

$$
\sum_{j=1}^n z_{1j}^{(k)} \left[ \log \frac{1}{\alpha_1^{(k)}} - \frac{x_j}{\alpha_1^{(k)}} \right] + \sum_{\ell=1}^L N_\ell z_{1\ell}^{(k)} \left[ \log \frac{1}{\alpha_1^{(k)}} - \frac{1}{\alpha_1^{(k)}} C^{(k-1)}_{1\ell} \right] 
+ \sum_{j=1}^n z_{2j}^{(k)} \left[ \log \frac{\beta_2^{(k)}}{\alpha_2^{(k)}} + (\beta_2^{(k)} - 1) \log \frac{x_j}{\alpha_2^{(k)}} - \left( \frac{x_j}{\alpha_2^{(k)}} \right)^{\beta_2^{(k)}} \right] 
+ \sum_{\ell=1}^L N_\ell z_{2\ell}^{(k)} \left\{ \log \frac{\beta_2^{(k)}}{\alpha_2^{(k)}} + (\beta_2^{(k)} - 1) \log \frac{\alpha_2^{(k-1)}}{\alpha_2^{(k)}} \right. 
+ \frac{\beta_2^{(k)} - 1}{\beta_2^{(k-1)}} \frac{e^{-\zeta_{\ell-1}} \log \zeta_{\ell-1} - e^{-\zeta_\ell} \log \zeta_\ell + \Gamma(0, \zeta_{\ell-1}) - \Gamma(0, \zeta_\ell)}{e^{-\zeta_{\ell-1}} - e^{-\zeta_\ell}} 
- \left( \frac{\alpha_2^{(k-1)}}{\alpha_2^{(k)}} \right)^{\beta_2^{(k)}} \Gamma \left( \frac{\beta_2^{(k)}}{\beta_2^{(k-1)}} + 1, \zeta_{\ell-1} \right) - \Gamma \left( \frac{\beta_2^{(k)}}{\beta_2^{(k-1)}} + 1, \zeta_\ell \right) \left\} 
$$

(A1)

where

$$
C^{(k-1)}_{1\ell} = \alpha_1^{(k-1)} + \frac{e^{-\xi_{\ell-1}} - e^{-\xi_\ell}}{e^{-\xi_{\ell-1}} - 1 - e^{-\xi_\ell}} \quad (A2)
$$

Note that in Eq.(A1) we have for the lower censoring interval boundary $\zeta_0 = 0$ that the term with $\ell = 1$ is well-behaved and reduces to

$$
\frac{e^{-\zeta_{\ell-1}} \log \zeta_{\ell-1} - e^{-\zeta_\ell} \log \zeta_\ell + \Gamma(0, \zeta_{\ell-1}) - \Gamma(0, \zeta_\ell)}{e^{-\zeta_{\ell-1}} - e^{-\zeta_\ell}} 
= -\gamma + e^{-\zeta_1} \log \zeta_1 + \Gamma(0, \zeta_1) 
$$

We need to maximise expression Eq.(A1) with respect to the parameters $\alpha_1^{(k)}, \alpha_2^{(k)}, \beta_2^{(k)}$, all other quantities being known from previous steps.

For the MLE equations the following expression arises during the lengthy computations:...
\[ D^{(k)}_{2\ell} = \frac{e^{-\xi_{\ell-1}} \log \xi_{\ell-1} - e^{-\xi_{\ell}} \log \xi_{\ell} + \Gamma(0, \xi_{\ell-1}) - \Gamma(0, \xi_{\ell})}{\beta^{(k-1)}_2} \]

\[ - \frac{1}{\beta^{(k-1)}_2} \left( \frac{\alpha^{(k-1)}_2}{\alpha^{(k)}_2} \right)^{\beta^{(k)}_2} \left[ \sum_{p=0}^{\infty} \frac{(-1)^p}{p!} \frac{\xi^{\beta^{(k)}_2 +1+p}}{\left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1 + p \right)^2} \right. \]

\[ - \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1 \right) \left( \log \xi_{\ell-1} - \log \xi_{\ell} \right) \]

\[ + \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1, \xi_{\ell-1} \right) \log \xi_{\ell-1} - \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1, \xi_{\ell} \right) \log \xi_{\ell} \]

\[ - \log \frac{\alpha^{(k-1)}_2}{\alpha^{(k)}_2} \left( \frac{\alpha^{(k-1)}_2}{\alpha^{(k)}_2} \right)^{\beta^{(k)}_2} \left\{ \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1, \xi_{\ell-1} \right) - \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1, \xi_{\ell} \right) \right\} \]

\[ (A3) \]

and for \( \ell = 1 \)

\[ D^{(k)}_{21} = -\gamma - e^{-\xi_1} \log \xi_1 - \Gamma(0, \xi_1) - \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} \]

\[ - \frac{1}{\beta^{(k-1)}_2} \left( \frac{\alpha^{(k-1)}_2}{\alpha^{(k)}_2} \right)^{\beta^{(k)}_2} \left[ - \sum_{p=0}^{\infty} \frac{(-1)^p}{p!} \frac{\xi^{\beta^{(k)}_2 +1+p}}{\left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1 + p \right)^2} \right. \]

\[ + \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1 \right) \log \xi_1 - \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1, \xi_1 \right) \log \xi_1 \]

\[ - \log \frac{\alpha^{(k-1)}_2}{\alpha^{(k)}_2} \left( \frac{\alpha^{(k-1)}_2}{\alpha^{(k)}_2} \right)^{\beta^{(k)}_2} \left\{ \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1 \right) - \Gamma \left( \frac{\beta^{(k)}_2}{\beta^{(k-1)}_2} + 1, \xi_1 \right) \right\} \]

\[ (A4) \]
A comparison to the direct maximisation of the terms in Eq. (8) as given in the Appendix A leads to comparable results and might justify our trick to speed-up convergence. Note that if we already knew the exact solution $\alpha_2^\infty$ and $\beta_2^\infty$, i.e. if we started with the true fixed point, these equations would be trivially satisfied.

We describe the censored EM-algorithm for a mixture of 1 exponential + 1 Weibull. The generalisation to arbitrary mixtures is obvious.

K. Burnham and D. Anderson, *Model Selection and Multimodel Inference - A Practical Information-Theoretic Approach* (Springer, New York, 2002) 2nd edition.

R. E. Kass and A. Raftery, Bayes factors, Journal of the American Statistical Association 90, 773 (1995).

E. Scalas, Mixtures of compound poisson processes as models of tick-by-tick financial data, Chaos, Solitons & Fractals 34 (1), 33 (2013).

G. Lubke and I. Campbell, Inference based on the best-fitting model can contribute to the replication crisis: Assessing model selection uncertainty using a bootstrap approach, Structural Equation Modeling: A Multidisciplinary Journal 23, 479 (2016).

B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Statist. 7, 1 (1979).

J. Loveless, Barbarians at the gateways, Commun. ACM 56, 42–49 (2013).

A. Feldmann, Characteristics of tcp connection arrivals, in *Self-Similar Network Traffic and Performance Evaluation* (John Wiley and Sons, Ltd, 2000) Chap. 15, pp. 367–399.

A. Arfeen, K. Pawlikowski, D. McNickle, and A. Willig, The role of the weibull distribution in modelling traffic in internet access and backbone core networks, Journal of Network and Computer Applications 141, 1 (2019).

N. Yannaros, Weibull renewal processes, Annals of the Institute of Statistical Mathematics 46 (4), 641 (1994).

L. Ponta, M. Trinh, M. Raberto, and E. Scalas, Modeling non-stationarities in high-frequency financial time series, Physica A 521, 173 (2019).