A single queue-reactive Hawkes model for the order flow

Peng Wu, Marcello Rambaldi, Jean-François Muzy, Emmanuel Bacry

Queue-reactive Hawkes models for the order flow

Peng Wu[1], Marcello Rambaldi [1], Jean-François Muzy[2], Emmanuel Bacry[1]
Université Paris-Dauphine and CEREMADE CNRS-UMR 7534
SPE UMR 6134, CNRS, Université de Corse, 20250 Corte, France

Abstract

In this work we introduce two variants of multivariate Hawkes models with an explicit dependency on various queue sizes aimed at modeling the stochastic time evolution of a limit order book. The models we propose thus integrate the influence of both the current book state and the past order flow. The first variant considers the flow of order arrivals at a specific price level as independent from the other one and describes this flow by adding a Hawkes component to the arrival rates provided by the continuous time Markov "Queue Reactive" model of Huang et al. [16]. Empirical calibration using Level-I order book data from Eurex future assets (Bund and DAX) show that the Hawkes term dramatically improves the pure "Queue-Reactive" model not only for the description of the order flow properties (as e.g. the statistics of inter-event times) but also with respect to the shape of the queue distributions. The second variant we introduce describes the joint dynamics of all events occurring at best bid and ask sides of some order book during a trading day. This model can be considered as a queue dependent extension of the multivariate Hawkes order-book model of Bacry et al. [4]. We provide an explicit way to calibrate this model either with a Maximum-Likelihood method or with a Least-Square approach. Empirical estimation from Bund and DAX level-I order book data allow us to recover the main features of Hawkes interactions uncovered in Bacry et al. in [4] but also to unveil their joint dependence on bid and ask queue sizes. We notably find that while the market order or mid-price changes rates can mainly be functions on the volume imbalance this is not the case for the arrival rate of limit or cancel orders. Our findings also allows us to clearly bring to light various features that distinguish small and large tick assets.

Keywords— Limit order book, market micro structure, Hawkes process, high frequency data, jump Markov process, ergodic properties, market simulator

1 Introduction

Building faithful models for the Limit Order Book (LOB) is a longstanding issue on which many efforts have been invested in the quantitative finance community. A rich literature of theoretical and empirical studies of limit order books has emerged in the last decade (see, e.g., [13] and [2] for a recent review). Modeling the LOB is a challenging task due to its intricate dependence structure. Indeed, the configuration of the limit order book is determined by the arrival of multiple types of orders: limit, cancel and market orders in the simplest setting, and the way these orders arrive on the market is non-trivial. For example, it is well known that order arrival inter-event times present strong and persistent autocorrelation (see e.g. [8]), implying that past order flow influences the current state of the book. At the same time, anecdotal as well as empirical evidence ([18]) suggests that market participants look at the state of the order book in order to take their trading decisions.

Models for the LOB can be roughly divided into two main classes. On one side are the models developed by the economics community where the focus is on the behavior of rational agents that act strategically to optimize their utility function (see e.g. [22]). Another stream of literature, beginning notably with [25], focus instead on the overall statistical properties of LOBs and assumes a certain simplified dynamics for the order flow in order to build mathematically tractable models that can reproduce at least partially some of these observed properties. This work contributes to the latter and builds on previous works in this field.

As stated above, in the pioneering work [25], the order book is seen as a purely stochastic system - a so called *zero intelligence* model - where orders arrive randomly, and where the interest is making testable predictions based on measurable inputs. [9] is one of the first paper to clearly frame the problem of LOB modeling in the context of queuing theory and Markov chains. By leveraging the properties of Markov chains, the authors are able to derive several conditional probabilities such as the likelihood of a mid-price

move or the probability of a limit order execution before a price change. The authors of [3] keep the same assumption of Poisson-driven independent queues and prove, using the theory of infinitesimal generators and Lyapunov stability criteria, the importance of the cancellation structure to ensure the stability of the LOB distribution and also show that under their model the price process converges to a Wiener process. Although the hypothesis made by these models are in disagreement with empirical facts, they present the advantage of being very tractable and to allow the derivation of many useful quantities analytically. In [1], the authors drop the assumption of uncorrelated order flow and introduce some memory effect by choosing to model the rate of limit and market order arrival, λ^L and λ^M by a Hawkes process ([15])

$$\lambda^\ell(t) = \mu^\ell + \sum_m \int \phi^{\ell m}(t-s) dN_s^m \ . \tag{1}$$

By setting the kernel function ϕ to an exponential form $\phi(t) = \alpha e^{-\beta t}$ the process (λ, N) has the Markov property and thus the authors are able to use a similar machinery to [3] in order to study the limiting behavior of their model. In [4] and [23], the authors also use multivariate Hawkes processes to analyze the order flow interaction at the fist level of the order book. Their model is calibrated without any assumption on the Hawkes kernel shapes using a non-parametric method.
In [16], the authors focus instead on the influence of the current state of the LOB on trading decisions. They propose a simple Markov model where the order flow arrival intensity depends only on the currently state of LOB through the available volume. They established conditions under which their model possess ergodic properties, making it possible to reproduce the empirical LOB queue size distributions as the invariant distribution of a Markov process. More recently, [19] extended the model of [16] by allowing the order book dynamics to depend also on the type of the order that led to a complete depletion of a level (i.e. a market or cancel order) and also by taking into account the order size. [19] thus depart slightly form the pure Markovian framework. They then discuss optimal market making strategies in the context of the model and also assess their performance on real data.
In this paper, we aim at contributing to this stream of literature by building on the work of [16] on one side and on the one of [1] and [4] on the other side by presenting a model where both dependence on past order flow and dependence on the state of the LOB are present. More precisely, we propose to consider stochastic LOB models that are multivariate Hawkes models whose parameters explicitly depend on the queue sizes of the order book. For a given set of event types ℓ that can occur in the considered LOB model, if $\vec{q}(t)$ represents the state of the book queues at some given time t, such a model could be simply written as the following generalization of the standard multivariate Hawkes model (Eq. (1)):

$$\lambda^\ell(t) = \mu^\ell\left(\vec{q}(t)\right) + \sum_m \int \phi^{\ell m}\left(t-s, \vec{q}(t)\right) dN_s^m \ . \tag{2}$$

where we have accounted for the possible explicit dependence on the queue sizes of both the exogenous intensities μ^ℓ and interaction kernels $\phi^{\ell m}$. We call this class of models *Queue Reactive Hawkes* (QRH) models. Let us notice that the issue of considering both self-excitation effects and dependence on some given state within a single model has also been considered very recently (during the completion of the present work) by Morariu-Patrichi and Pakkanen [21]. These authors proposed a general framework called "state dependent Hawkes process" where the Hawkes kernels ϕ are functions of some state process $X(t)$ that can take a finite number of values and that switches from one state to another one when an event of the Hawkes process occurs and according to a transition rule that depends on the type of this event. The specific application of this framework to LOB modeling proposed in [21] mainly consists in considering either the volume imbalance or the spread as the state variable. Let us mention another very recent and related work in the paper of Daw and Pender [11] that defined and studied a Markov process constructed a pair of inter-dependent processes (N_t, Q_t), where N_t is a counting process and Q_t a queuing process.
In this paper our purpose is twofold: We first investigate in which respect adding Hawkes self- and cross-excitation properties to the "Queue Reactive" model of Huang et al. [16], may improve this model. To achieve this goal, we consider a simple version, referred to as QRH-I, of the general model described by (2) where the Hawkes kernels do not depend on the queue sizes and the various queues are considered as independent. In a second part, we consider a queue state dependent version of level-I LOB Hawkes model introduced by Bacry et al. [4] where we account for bid and ask queue interactions and we suppose that both exogenous intensity and interaction kernel matrix share the same multiplicative queue dependencies. We call this model the QRH-II order book model. By calibrating these models using high frequency data from Eurex future markets, we show that both models achieve a better fit of the data than their pure

queue reactive or Hawkes restrictions. Let us emphasize that, like the QR model of [16], the QRH-I model can be considered as a model for both the order flow and the queue state whereas, within the QRH-II model, our main concern will be to improve the order flow description, in particular we will consider the queue as an exogenous imput.
The rest of paper is organized as follows. In Section 2 we elaborate on the QRH-I model, i.e. on a single-queue model that consists in adding an order flow dependence as provided by a multivariate Hawkes process to the Queue Reactive (QR) model introduced in [16]. We show how such a model can be calibrated using a maximum likelihood approach and prove that, very much like the QR model, under some reasonable assumption, the queue size admits an invariant distribution. We then compare the likelihood of QRH-I model with both a standard Hawkes model with no state-dependence and with the model of [16] on real data form the Eurex exchange. Comparisons with empirical data show that QRH-I model represents an important improvement of the QR model not only with respect to the inter-event time statistics buy also regarding the predicted shape of the equilibrium queue size distribution. In Section 3, we start instead from the point of view of level-I book model of [4], that is a multivariate Hawkes model for all events occurring at the first level of the order book. The QRH-II model is defined by considering a multiplicative dependence of Hawkes kernel matrix and exogenous intensities on both the best bid and best ask queue states. We show that such a state dependence is a very meaningful assumption. The empirical results provided by the model calibration using Eurex future data are then discussed. Concluding remarks and prospects for future research are provided in Section 4. Technical results like the proof of the ergodicity of QRH-I, the model calibration issues by Maximum Likelihood or Least-Square approaches are given in the Appendices.

2 A Queue Reactive Hawkes model for a fixed-price best limit

2.1 Adding memory to the Queue Reactive model of Huang et al.

As mentioned in the introduction, Huang, Lehalle and Rosenbaum present, in [16], a model where the order flow arrival at a given price level is modeled as an inhomogeneous Poisson process with an intensity that depends only on the current state of the order book and in particular on the queue sizes. They name this property *Queue Reactive* (QR). Let us briefly recall here the main lines of the QR approach which will be shared by our model.
The order book is seen as a $2K$-dimensional vector, where K represents the number of available levels on each side, the prices living on a grid whose resolution is the tick size *tick*, i.e., the unit of price variations. The bid and ask sides are separated by a reference price p_{ref} which is equal to the midprice p_{mid} (i.e., the mean value of best bid and best ask prices) if the spread is an odd multiple of the tick size and equal to $p_{\text{mid}} \pm \frac{tick}{2}$, whichever is closer to the previous p_{ref}, if the spread is an even multiple of the tick size. The price levels on the bid side are denoted as $\{Q_{-i}\}_{i=1\dots K}$, those at the ask side as $\{Q_i\}_{i=1\dots K}$ while the quantities available at these levels by $q_{\pm i}$. The queue sizes are modified by the arrival of limit, market and cancel orders. For the sake of simplicity, all orders are assumed of unitary volume, so a limit order adds one unit to the queue, while a market or a cancel order subtract one unit. We will denote by λ_i^L, λ_i^M and λ_i^C the arrival intensities of respectively limit, market and cancel orders on the queue i. In the simplest version of the QR model of Huang et al., all the queue sizes are independent one from each other and, for some given queue i, the intensity λ_i^L, λ_i^M and λ_i^C depends only on the queue size q_i, namely:

$$\begin{aligned} \lambda_i^L(t) &= \mu_i^L(q_i(t^-)) \\ \lambda_i^C(t) &= \mu_i^C(q_i(t^-)) \\ \lambda_i^M(t) &= \mu_i^M(q_i(t^-)) \end{aligned} \tag{3}$$

where the functions $\{\mu_i^\ell(q)\}_{i,\ell}$ are the parameters of the model. They correspond to the rates of a birth-death Markov process and can easily be estimated via maximum likelihood which, in this case, amounts to the computation of conditional empirical means of intensities defined in [16]. As the labeling of a price level is relative to the reference price, when p_{ref} changes, the level labels also change. Hence, the estimation is performed on intervals where p_{ref} is constant and each period is regarded as an independent realization of the process. As shown in [16], the QR model and its extension accounting for the queue interactions is an ergodic continuous time jump Markov process provided the limit order rate is bounded for large queue sizes, and the rate at which orders are removed is larger than the rate which increases the queues. In that respect, the QR model represents a simple and parsimonious Markov model that

allows one to account for the state dependent nature of the book dynamics and that is able to describe the (stationary) distribution of the queue sizes.
Our goal is to consider an extension of Huang *et al.* QR model with independent queues that is able to account for both queue reactive and for the memory effects in the order flow. This can be done by combining the dependence on the current state of LOB with the one on past order flow events as given by a multivariate Hawkes process within a QRH model. Note that, as discussed in the introduction (Eq. (2)), in principle both Hawkes kernels and the baseline intensities could depend on the LOB state and indeed we will explore such a possibility in the next section. Here however we are interested in the simplest modification of the model of Huang *et al.* that accounts for the dependence on the past order flow. In that respect our model introduces state-dependence in the multivariate Hawkes framework by allowing exclusively exogenous intensities to depend on the queue size. Furthermore, since our database mainly concerns the best bid and ask, we consider only the positions $Q_{\pm 1}$. By bid-ask symmetry (supported by empirical observations), both processes can be assumed to have the same law and we will henceforth drop the subscript i. Accordingly, N_t^L, N_t^C, N_t^M, $\lambda^L(t)$, $\lambda^C(t)$ and $\lambda^M(t)$ will denote the counting processes and their associated intensities defined by the arrivals of respectively limit, cancel and market orders at best bid (or alternatively at best ask). We refer to this variant of the QRH model as the QRH-I model since it mainly concerns a single LOB queue (at best bid or best ask). For $\ell, m \in \{L, M, C\}$ (for respectively limit, market and cancel orders), the QRH-I model thus defines $\lambda^\ell(t)$ as:

$$\lambda^\ell(t) = \mu^\ell(q(t^-)) + \sum_m \int_0^t \phi^{\ell m}(t-s) dN_s^m \ . \tag{4}$$

where the queue size $q(t)$ is simply $q(t) = q(0) + N_t^L - N_t^M - N_t^C$. Since market and cancel orders can only be sent when the queue is non-empty, in the case when $q(t) = 0$, the previous expression should be replaced by $\lambda^\ell(t) = 0$ for $\ell = M, C$. The baseline intensities $\{\mu^\ell(q)\}$ depend on the LOB state q while the Hawkes kernels $\phi^{\ell m}(t)$ account for the effect of past orders of type m occurrence on the current intensity $\lambda^\ell(t)$. In full rigor, to complete the model definition, one should specify the law of the the initial queue size $q(0)$. Since, has shown below, we will consider a situation when the queue process is a component of an ergodic vector Markov process, the choice of this law is not pertinent an we simply choose $q(0) = 0$.
Let us note that the intensity function of the QR model (with independent queues) and the one of a standard Hawkes model could both be treated as a special case of the intensity function of QRH-I model. One recovers the QR model when the Hawkes kernels are zero and a standard multivariate Hawkes model for constant baseline intensities. We proceed as in [16] and we assume that the model (4) holds in periods when the reference price is constant, and furthermore that such periods can be considered as independent realizations. Note that by doing so we reset the Hawkes memory every time there is a change in the reference price. We will discuss this point below when analyzing the empirical results and we will drop this assumption in the model we present in Section 3.
The model parameters can be estimated using the maximum likelihood method. The log-likelihood L of a D-dimensional point process where the components do not share any parameters has the following general form (see [10], page 21)

$$L(\theta) = \sum_{\ell=1}^{D} L_\ell(\theta), \quad \text{with} \quad L_\ell(\theta) = \int_0^T \log \lambda^\ell(t;\theta|\mathscr{F}_t) dN_t^\ell - \int_0^T \lambda^\ell(t;\theta|\mathscr{F}_t) dt \tag{5}$$

where θ denotes the parameter set and λ^ℓ is the intensity function of the ℓ-th component.
To use the method in practice, a parametric form must be specified for the interaction kernels $\phi^{\ell m}$ in Eq. (4). A standard choice is to consider that $\phi^{\ell m}$ can be written as sum of exponential kernels:

$$\phi^{\ell m}(t) = \sum_{u=1}^{U} \alpha_u^{\ell m} \beta_u e^{-\beta_u (t-s)} \tag{6}$$

where $\alpha_u^{\ell m}$ are parameters of the model and β_u, U are suitably chosen hyper-parameters. This choice also presents the important advantage that the resulting log-likelihood is a convex function of the model parameters. To faciliate the notation, we use $\vec{\mu}$ and $\vec{\alpha}$ to represent all μ^ℓ and $\alpha_u^{\ell m}$. With such parametriza-

tion, $\theta = (\vec{\mu}, \vec{\alpha})$.The log-likelihood of the QRH-I model (4) thus reads:

$$\begin{aligned} L(\theta) = &\sum_{\ell=1}^{D}\sum_{k=1}^{N^\ell} \log\left(\mu^\ell(q(t_k)) + \sum_{m=1}^{D}\sum_{u=1}^{U} \alpha_u^{\ell m}\beta_u \int_0^t e^{-\beta_u(t-s)} dN_s^m\right) \\ &- \sum_{\ell=1}^{D} \int_0^T \left(\mu^\ell(q(s)) + \sum_{m=1}^{D}\sum_{u=1}^{U} \alpha_u^{\ell m}\beta_u \int_0^s e^{-\beta_u(s-v)} dN_v^m\right) ds \end{aligned} \tag{7}$$

As we show in Appendix A.2, the specific choice of a sum of exponential functions (Eq. (6)) allows for a computationally efficient calculation of the log-likelihood and of its gradient.
Another important advantage of the parametrization (6) is that it allows us to work within the framework of Markov processes. Let us note $I = \{C, M\}$ the set of order types that decrease the queue size and $J = \{L\}$ the set of order types that increase the queue size. The queue size $q(t)$ thus simply corresponds to:

$$q(t) = \sum_{m\in J} N_t^m - \sum_{\ell\in I} N_t^\ell \, . \tag{8}$$

If one defines

$$o_{\ell mu}(t) = \int_0^t \alpha_u^{\ell m} e^{-\beta_u(t-s)} dN_s^m \, , \tag{9}$$

and denote by $\vec{o}(t)$ the vector obtained by a vertical stacking of $o_{\ell mu}(t)$ ($u \in \{1, \dots, U\}$, $\ell, m \in \{L, M, C\}$), then $\left(\begin{smallmatrix} q(t) \\ \vec{o}(t) \end{smallmatrix}\right)$ is vector Markov process. This property can be proved exactly along the same lines as in Proposition 2.2 of [17]. Moreover, if one assumes that

$$\sum_{\ell\in I} \mu^\ell(q) \geq c_\ell q \text{ and } \mu^m(q) \leq c^*, \forall m \in J \, ,$$

we show, in Appendix A.1, with the help of Lyapunov functions approach, that the process $\left(\begin{smallmatrix} q(t) \\ \vec{o}(t) \end{smallmatrix}\right)$ is V-uniformly ergodic which notably means that $q(t)$ admits an invariant distribution and that this equilibrium is reached exponentially fast.

2.2 Calibration results

Data In this study we use tick by tick level L1 data of Bund future and DAX index future traded on the Eurex electronic future market. The data span the period from October 1st 2013 to September 30th 2014. The dataset consists in snapshots of the first level of the order book, each with a timestamp indicating the record time with microsecond precision, that provide prices and outstanding quantities. Every time a trade occurs a specific line is added to the dataset, thus allowing to precisely determine the type of order (i.e. limit order, cancellation, or market order[1]) that lead to a change in the LOB. The Eurex future market is open from 8 a.m. to 10 p.m., Frankfurt time, however thorough this paper we only consider the time slot from 9 a.m. to 9 p.m. in order to capture the most active period. In Table 1, we report some descriptive statistics of our datasets. We note that, from the micro-structural point of view, the Bund future can be considered as a large tick asset, with and average spread very close to one, while the DAX has a considerably smaller perceived tick size compared to the Bund. This difference at the market microstructure level will reflect also in the queue dynamics and in the result of our model. Indeed, the queues on the Bund are often large as the midprice stays constant for relatively long periods of time while bids/offers accumulate at the best quotes, while on the other hand on the DAX the midprice changes more frequently thus resulting in slimmer queues at the best quotes. Further details on the datasets as well as a more detailed description of the inter-event time distribution can be found in [23].
Let us emphasize that since in our setting the order book is seen as a collection of independent queues, we impose *a priori* a strict bid-ask symmetry and all the results presented in Sec. 2 correspond to averaging estimations obtained on the bid and ask sides.

[1] With a slight abuse of language, in this paper we use the term "market order" to denote any order that immediately gives rise to a trade, regardless to whether or not it has a limit price.

	# L	# C	# M	Avg. spread	Med. spread	AES	Med. inter-event time
Bund	5.41×10^7	4.67×10^7	6.29×10^6	1.012	1.0	6.34	4.89×10^{-4}
DAX	5.46×10^6	5.62×10^6	6.68×10^5	1.591	2.0	1.30	1.73×10^{-3}

Table 1: Descriptive statistics of our dataset. Average number of limit, cancel and market order at the best quotes per day. Average and median spread (measured in ticks) and average order sizes expressed in contracts.

Estimation and goodness-of-fit analysis In order to estimate the parameters of our model, for each day in our sample we first compute the reference price p_{ref} as specified at the beginning of this section. Then we determine the queue sizes, as in [16], we assume that orders sizes for all events is a constant corresponding to the average event size (AES), defined as the average volume of all types of orders arriving at $Q_{\pm 1}$. We therefore measure the queue q in units of AES as

$$q(t) = \lceil \frac{v(t)}{\text{AES}} \rceil \tag{10}$$

where $\lceil \rceil$ is the ceiling function and $v(t)$ is the volume available at time t in the queue. In Figure 1 we show the empirical distribution of the so-defined $q(t)$ for Bund (left panel) and DAX (right panel). These distributions are obtained by sampling the book state every 30s over the whole time period. Notice that the state $q(t)$ is set to 0 if and only if $v(t) = 0$. We observe that the Bund future presents a broader and smoother distribution as compared to the one of the DAX, which is on the contrary more concentrated on small queue sizes. This is a direct consequence of the different perceived tick sizes as we observed above.

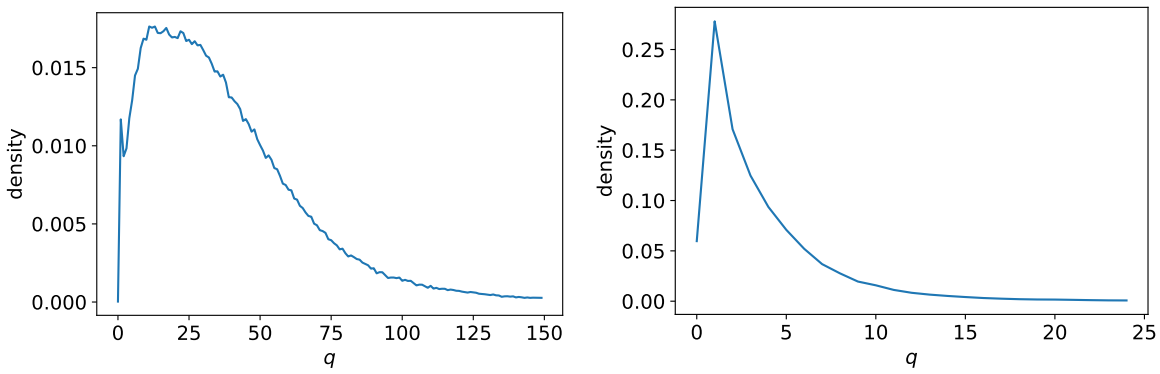

Figure 1: Empirical distribution of the queue states (measured in unit of AES) as defined in (10). Left: Bund future. Right: DAX future.

Once p_{ref}, $Q_{\pm 1}$, and $q_{\pm 1}$ are determined, we divide each day in periods where p_{ref} is constant. Then, each period is considered as an independent sample and we determine the parameters of our model by numerically optimizing the joint log-likelihood over all the so-obtained independent samples. In total, we have $1,207,099$ periods for the Bund and $417,581$ for the DAX. Notice that statistical estimation purpose, in the results reported below we considered only samples that contain at least a total of 20 events and disregarded the other ones. Notice that if s_k stands for the length of the realization k and k_t the index of the realization located around time t, the quantity

$$\tau_m = \frac{\mathbb{E}\left[s_k^2\right]}{\mathbb{E}\left[s_k\right]}$$

represents the average length of the realization k_t if one chooses t at random (i.e. with a uniform probability). This quantity is pertinent when performing averages over a fixed grid of times $\{t_j\}_j$. For the Bund we have $\tau_m \simeq 100\ s$ while for the DAX we estimated $\tau_m \simeq 16\ s$.
As we pointed out above, in order to estimate our model we need to fix the number of exponential decays U as well as the values of the decays β themselves. We take $U = 3$ and we set $\beta_1 = 60s^{-1}$, $\beta_2 = 1500s^{-1}$, $\beta_3 = 5500s^{-1}$ for the Bund and $\beta_1 = 40s^{-1}$, $\beta_2 = 2100s^{-1}$, $\beta_3 = 5200s^{-1}$ for the DAX. We experimented

Bund				
	L	AIC	BIC	# parameters
QR	2.046×10^7	-4.093×10^7	-4.092×10^7	450
QRH	2.055×10^8	-4.110×10^8	-4.110×10^8	477
DAX				
	L	AIC	BIC	# parameters
QR	7.268×10^5	-1.453×10^6	-1.452×10^6	75
QRH	9.506×10^6	-1.901×10^7	-1.901×10^7	102

Table 2: Log-likelihood, AIC, and BIC values for the three considered models for Bund and DAX data.

Bund			
	Difference of log-likelihood	df	p-value
$H_0 =$ QR, $H_1 =$ QRH	$3.7 \cdot 10^8$	27	$< 10^{-16}$
DAX			
	Difference of log-likelihood	df	p-value
$H_0 =$ QR, $H_1 =$ QRH	$1.8 \cdot 10^7$	27	$< 10^{-16}$

Table 3: Likelihood ratio test statistic and p-values for the case where the null hypothesis is the QR model and for the case where the null hypothesis is a standard Hawkes model. The "degree of freedom" ("df") value indicates the difference in the number of parameters between the two models.

with several combinations of U and β and we found these ones to represent a good compromise between the total number of parameters to estimate (the number of parameters α grows linearly with U) and the model goodness of fit as measured by (penalized) log-likelihood.
The maximum likelihood allows us to perform a quantitative comparison of the QR and QRH-I models in terms of goodness of fit, which is one of the central results of this paper. For completeness we also consider a standard Hawkes model, i.e. with no dependence on the queue state. In Table 2 we report the log-likelihood values for the three models as well as the Akaike Information Criterion (AIC) score

$$\text{AIC} = 2k - 2L \tag{11}$$

and the Schwartz information criterion (BIC) score

$$\text{BIC} = k \log N - 2L \tag{12}$$

where k is the number of parameters, L is the log-likelihood and N is the total sample size (number of events in our case). These scores allow one to compare nested models by their likelihood while taking into account the different number of parameters (lower score is better). By looking at the values reported in Table 2, we observe that the QRH-I model has better scores in terms of AIC and BIC for both assets. We can also use the likelihood ratio test in order to compare the models. Indeed the QRH-I model reduces to the QR model when all the α are set tot zero. Likewise, the QRH-I model reduces to a standard Hawkes model when, $\forall \ell$, $\mu^\ell(q) = \mu^\ell$, i.e the dependence on the queue state is dropped. We report the test statistics

$$\text{LR} = 2(L(\hat{\theta}_1) - L(\hat{\theta}_0)) \tag{13}$$

where $\hat{\theta}_1$ and $\hat{\theta}_0$ are the maximum likelihood estimates for the null and for the alternative model respectively, and p-values for the likelihood ratio test in Table 3. We note that both the QR and the Hawkes model are rejected with a very high degree of significance when compared to the QRH-I model.
To complete the goodness-of-fit comparison of the models, we look at the inter-event time distribution. In particular, in Figure 2 we compare by means of a quantile-quantile plot the empirical inter-event times distribution with the ones produced by simulations of the calibrated QR and QRH-I models. It is clear from the figure that the QRH-I model reproduces strikingly better the empirical inter-event distribution, indicating that including the dependence on the past event is crucial in order to build a faithful model for the order flow fluctuations.

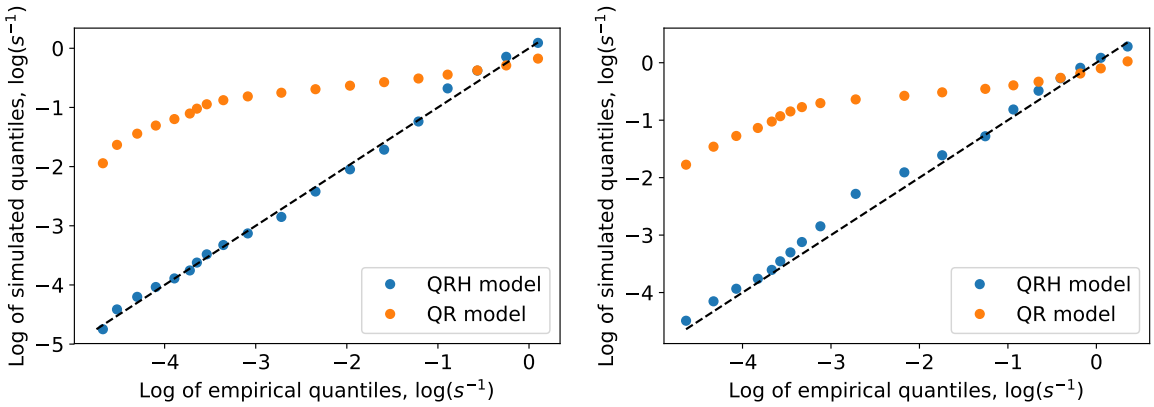

Figure 2: Log qq-plot of inter-event times. Log of quantiles of inter-events times simulated by model (horizontal) is plotted against log of empirical quantiles (vertical). Left: Bund future. Right: DAX future.

The results presented in this section suggest that both LOB-state dependence and memory effects due to correlation in the order flow are relevant variables that need to be taken into account in order to build a faithful model for the order book dynamics. Crucially, adding a order flow dependence in the form of a Hawkes term dramatically increases the model likelihood as well as its capability of reproducing the observed inter-event time distribution.

State dependency and Hawkes matrix empirical estimations In Figure 3 we report the estimated parameters $\mu(q)$ for the QR model, while in Figure 4 we plot the analogous quantities for the QRH-I model (4). We can make two general remarks while comparing these plots. First, we note that the dependence on the queue size captured by the two models are roughly concordant, in that functions $\mu^\ell(q)$ have similar shapes in both models. However let us remark that the values estimated within the QRH-I model are much smaller, indicating that a large part of the intensity is now explained by the self- and cross-exciting Hawkes components (see the discussion below). One can notice some differences between the Bund and DAX results, most likely stemming from the different order book dynamics typical of large and small tick assets respectively. As in [16], we observe a decreasing rate of Market orders arrivals as the queue size increases. This can be explained by the fact that agents tend to consume liquidity faster as this liquidity becomes rare. We also find that, when, $q(t)$ is large enough, the rate of cancellation is an increasing function of the queue size. This is an expected feature assumed in most former LOB models (see e.g., [25, 9]), since cancellations are more likely to occur when they are many active limit orders. As shown in Appendix A.1, this behavior ensures the ergodicity of the queue process. Let us finally notice, that unlike the observed behavior in [16] on specific stocks, we don't observe that the intensity of limit order insertion is almost independent of the queue size. It is rather a decreasing function of the queue size probably reflecting a lesser quest for priority when q is large.
It is also interesting to look at the quantity

$$e^\ell(q) = 1 - \frac{\mu^\ell(q)}{\Lambda^\ell(q)} \tag{14}$$

where

$$\Lambda^\ell(q) = \mathbb{E}\left[\lambda^\ell(t) | q(t^-) = q\right] \tag{15}$$

is the average intensity in a given state q. $e^\ell(q)$ corresponds to the fraction of the total average intensity explained by the endogenous self- and cross-exciting mechanism as a function of the queue size q, While $\Lambda^\ell(q)$, in the case of the QR model, is directly provided by the parameter $\mu^\ell(q)$, for the QRH-I model it is given by the contribution of both the baseline intensity $\mu^\ell(q)$ and the Hawkes interactions. Unlike standard multivariate Hawkes processes, the QRH-I model does not admit a closed form formula of $\Lambda^\ell(q)$ from its parameters. Therefore, while $e^\ell(q)$ is trivially zero for the QR model, we resort to numerical computation of $\Lambda^\ell(q)$ in order to compute $e^\ell(q)$ for the QRH-I model.
The result are shown in Figure 5, where we have plotted the estimated $e^\ell(q)$ for all types of orders and for both the Bund (top panels) and the DAX (bottom panels) futures. Overall we see that a large part, from 60% to 80%, of the total average intensity is explained by the self- and cross-exciting effect. We

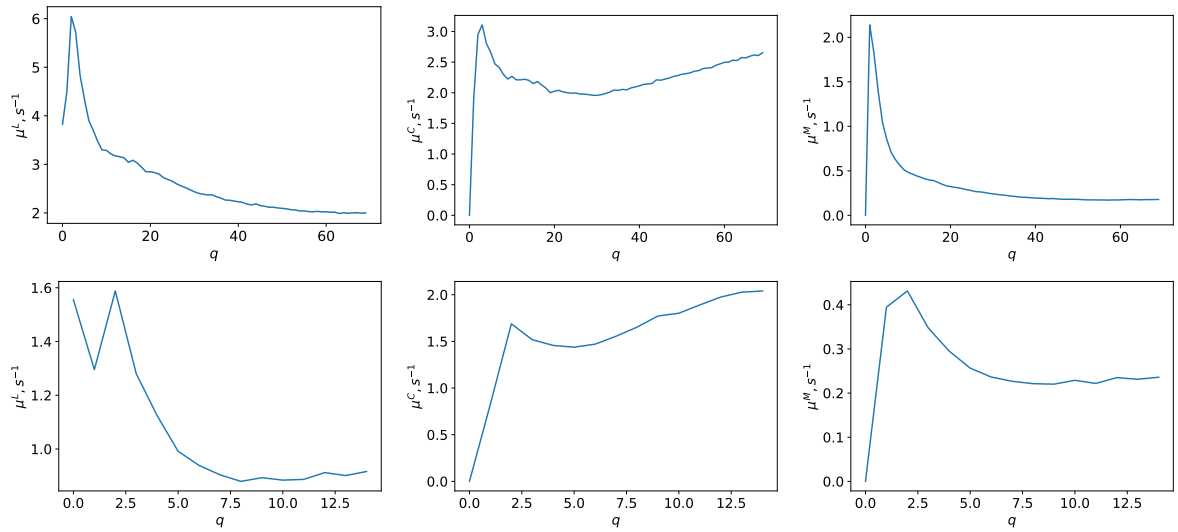

Figure 3: From left to right: estimated values μ_q for limit order insertion, limit order cancellation and market orders, QR model. Top row: Bund future. Bottom row: DAX future.

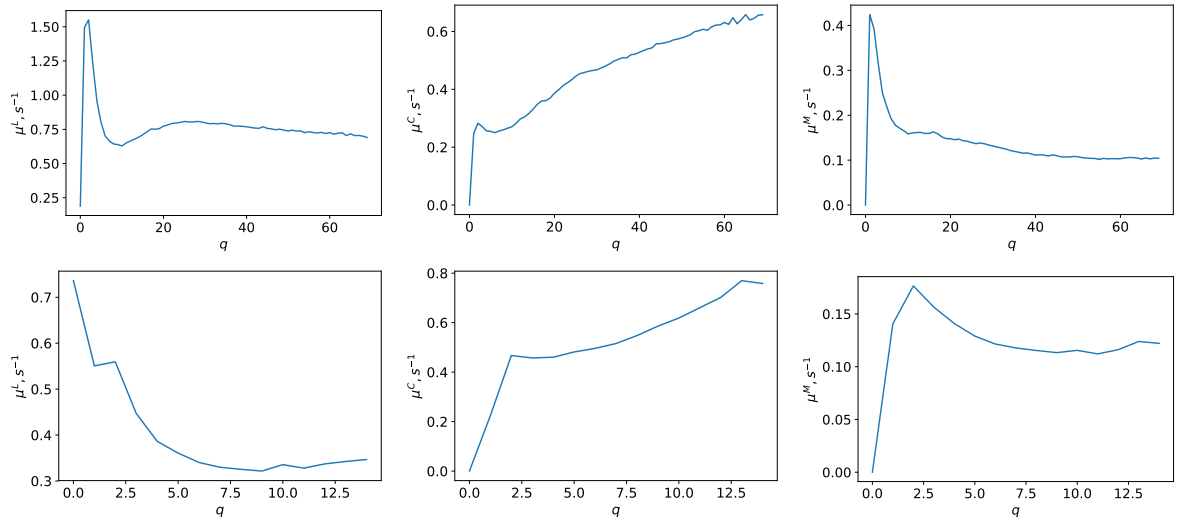

Figure 4: From left to right: estimated values μ_q for limit order insertion, limit order cancellation and market orders, QRH model. Top row: Bund future. Bottom row: DAX future.

note that for cancel and market orders, the intensity is maximally explained by the Hawkes term when the queue is small. This is likely to result from the persistence in the order flow and in particular of the relevance of the self-exciting term as we noted in Figure 6, which in case of market and cancel orders tends to empty the queue. This explanation is corroborated by the observation that the opposite effect is found for limit orders, namely an higher endogeneity for higher values of q.
To complete the analysis of the QRH model results, in Figure 6 we plot in a color map the Hawkes kernel norms $|\phi^{\ell m}| = \int_0^\infty \phi^{\ell m}(t)dt$. Note that in our setting these quantities are simply given by $|\phi^{\ell m}| = \sum_{u=1}^{U} \alpha_u^{\ell m}$. As discussed in [5], these quantities represent the average direct effect of an event of type m (columns) over the intensity of type ℓ (rows) events. Hawkes kernel matrices of order book events have been extensively studied in [4, 23]. Here, we note that despite the addition of the queue-dependent term, we recover many of the features already observed in previous studies, such as the strong diagonal component corresponding to self-excitation, likely the result of correlation in the order flow induced by order splitting strategies. We also confirm that market orders influence liquidity much more than the opposite effect. In particular, since here we look at interaction on the same side of the book, we note that market order have on average an exciting effect on cancellations. As observed in the aforementioned studies, a flow of market order at a given price signals that the "true" price is closer to that side and

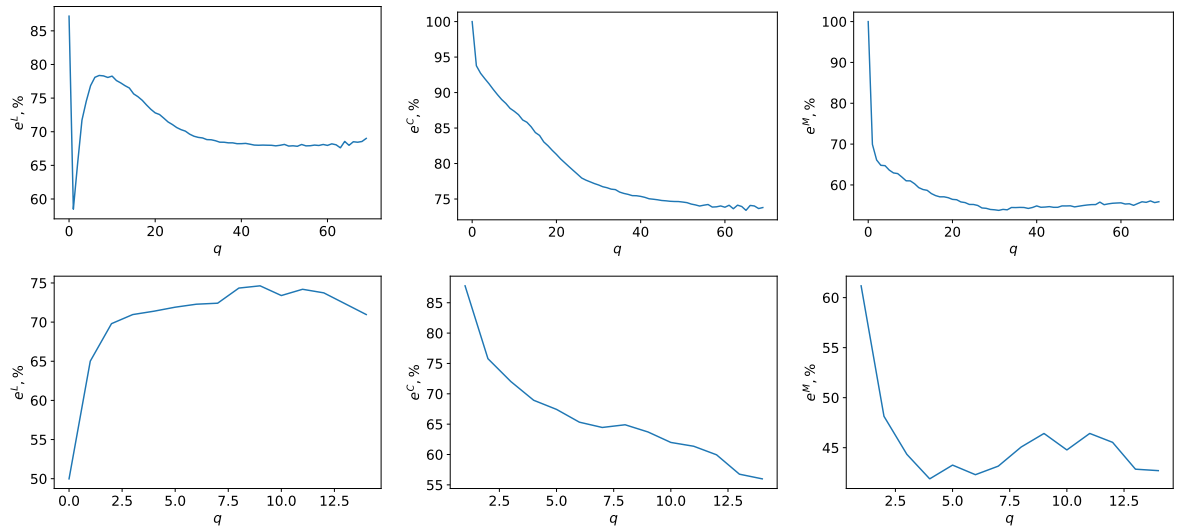

Figure 5: From left to right: Endogenous fraction for limit order insertion, $e^{\ell}(q)$ as defined in Eq. (14)), for limit ($\ell = L$), cancellation ($\ell = C$) and market ($\ell = M$) orders by QRH model. Top row: Bund future. Bottom row: DAX future.

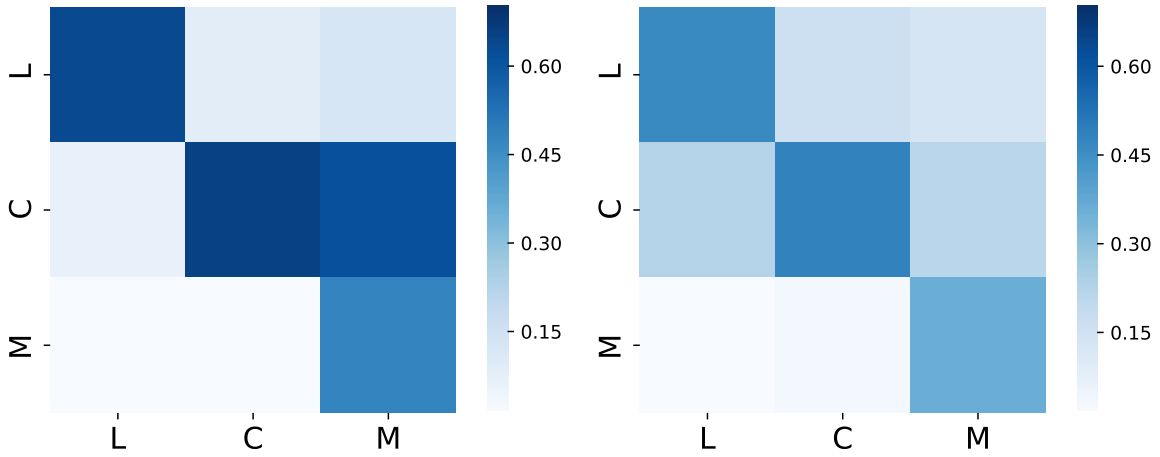

Figure 6: Kernel norm matrices for Bund future (left) and DAX future (right).

therefore liquidity adapts, with outstanding orders being canceled in order not to be adversely selected.

Equilibrium and empirical queue size distributions In Ref. [16], the authors emphasized that the QR model provides a simple framework to account for the observed queue size distributions in the order book. For that purpose they have shown that the model invariant distribution fits quite well the empirical laws notably at the first bid/ask levels. In Appendix A.1, we show that, under some conditions that appear to be empirically fulfilled, the QRH-I model is also an ergodic process and the queue size can thus be described by its invariant distribution. Before comparing the performances of QR and QRH-I models with respect to their prediction of the equilibrium queue size distribution, let us emphasize that some caution is needed when addressing this issue. Indeed, this distribution, even if reached exponentially fast, does necessarily correspond to the empirically observed queue law since when the queue is empty, the reference price has a non-vanishing probability to change. This directly implies that for small values of the queue size, the invariant distribution is not supposed to account for the observed values from snapshots of the empirical book state. Moreover, since the initial queue size has no reason to be drawn with the invariant distribution, this law is pertinent only after a short delay that has to be compared to the length of each realization, i.e., the time period between two changes of reference price. The exponential rate involved in the ergodic theorem is however hard to estimate, one can use a proxy as given for example by the exponential decay of empirical queue size autocorrelation function. That is an alternative measure of

a mixing coefficient that can be, under some conditions, related to the distribution relaxation time [7]. If one assumes that the decay of autocorrelation of the queue size takes the form $\rho(t) = a \exp^{-t/\tau_c}$, we find empirically that $\tau_c \simeq 15\ s$ for the Bund future and $\tau_c \simeq 2\ s$ for the DAX. For both assets these correlation characteristic scales have to be compared with the average realization length, namely $\tau_m \simeq 100$ s for the Bund and $\tau_m \simeq 16s$ for the DAX. Since in both cases we have $\tau_c \ll \tau_m$, it is likely that the invariant distribution is pertinent to account for the queue size distribution as observed at randomly chosen times. With previous observations in mind, we now proceed at the comparison of the empirical queue distribution, measured by taking snapshots of the book every 30s and the invariant distributions produced by the QR and QRH-I models. Since we do not have any explicit formula for the QRH-I model, we estimate the invariant distribution of $q(t)$ by performing a simulation over a long time period. The invariant measure of the the QR model can be directly deduced from the estimations of $\mu^\ell(q)$ in Figure 3 using the analytical formula in Sec. 2.3.3 of Ref. [16]. The plots the both QR and QRH-I invariant measures together with the empirical queue size distribution for both Bund and DAX are reported in Figure 7. First of all, we observe that the QRH-I model provides in both cases of better fit of empirical data, notably in the tail region, than the QR model. The latter is particularly far from the observed distribution in the large tick case of the Bund data. Its performance for the smaller tick asset (DAX) are closer to the results reported in [16] for stock data. Beyond the fact that this striking difference between large and small tick assets is hard to explain (though the analytical formula in [16] shows that the overall shape of the distribution can vary quite drastically when on changes the respective behavior of the $\mu^\ell(q)$ functions) our findings show that accounting for the Hawkes self-interaction within a queue reactive model is important not only to describe correctly the order flow dynamics but also provides a better model for the queue size distributions.

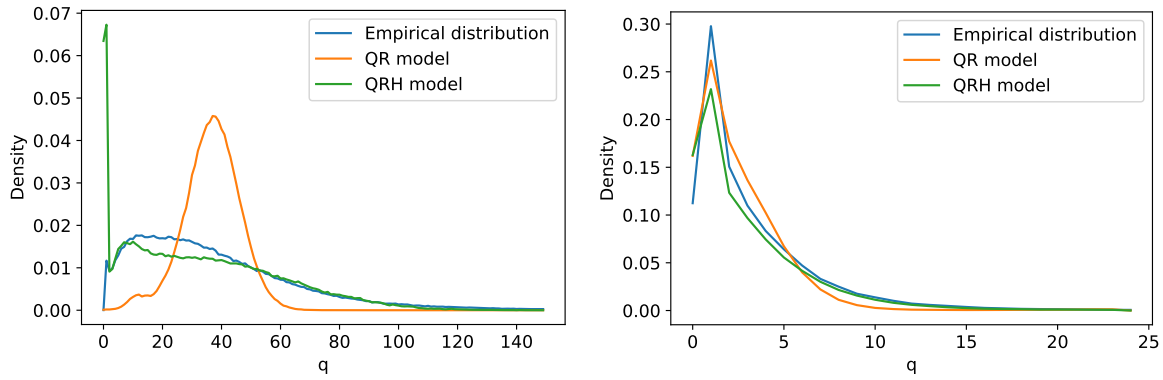

Figure 7: Comparison of the invariant distributions of the QR and QRH models with the empirical one. Left: Bund future. Right: DAX future.

We notice furthermore that the distribution of queue size simulated by QRH-I model deviates from the empirical distribution especially around states when the queue is small. As discussed previously, when the queue is empty, the reference price has a large probability to change and therefore the corresponding empirical sample to stop. This results in a statistical bias with respect to the "theoretical model" where such event type does not exist since when the queues empty nothing happens until a new limit order arrives. The states with a low queue values are therefore more likely to be visited in the model than in empirical observations.

3 A Queue Reactive Hawkes model for the best limits of the orderbook

In the previous Section, we presented a model for a single best limit with a fixed price, i.e., the model is reset when the price of the best limit changes. Thus, it does not account for the cross dynamic between the two best limits (best bid/best ask) nor it accounts for the changes of their corresponding price. This Section is devoted to a model that tackles these drawbacks. Both best limits are modeled along with mid-price changes. Hence, contrarily to the previous model (and also to the models presented in [16]), resetting of the model only occurs at market close time. This new model is built by adding a queue-dependency to the model of Bacry et al. [4]. The model of the previous section was referred to as QRH-I

(one-side LOB modeling), this new model will be referred to as QRH-II (two-side LOB modeling).

3.1 The QRH-II model : Adding queue-dependence to Bacry et al.'s model

In [4] the authors consider eight event types at the best level of a LOB, namely

P^+ (P^-) for events that moves the midprice[2] up (down) independently on the size of this move,

L^a (L^b) for limit orders at the best ask (bid) that do not change the midprice,

C^a (C^b) for cancellations at the best ask (bid) that do not change the midprice,

M^a (M^b) for market orders at the best ask (bid) that do not change the midprice,

where, for any $\ell \in \{P^+, P^-, L^a, L^b, C^a, C^b, M^a, M^b\}$, the best ask (resp. bid) level refers to the first non empty level on the ask (resp. bid) side. Let us define N_t^ℓ as the counting process associated with events to type ℓ and $\lambda^\ell(t)$ the associated conditional intensity. Let us point out that that this description of the first limits of the LOB is significantly different from the approach taken in the previous section or in [16] in which the best limit queue size can be 0. Also, this model no longer needs to be reset regularly and does not use any notion of reference price p_{ref}.
The authors in [4] consider a multivariate Hawkes model where each event type can influence, and be influenced by, the others so that the conditional intensities read:

$$\lambda^\ell(t) = \mu^\ell + \sum_m \int_0^t \phi^{\ell m}(t-s) dN_s^m, \tag{16}$$

where ℓ and m can take any value among the 8 values $\{P^+, P^-, L^a, L^b, C^a, C^b, M^a, M^b\}$. In their work the kernels ϕ are estimated using the non parametric estimation method first described in [6]. This model allows the authors to highlight the rich influence structure between events in a limit order book, including the high-frequency midprice reversion and the persistent autocorrelation in the order flow determined by order splitting strategies but also some more refined market-maker induced dynamics (see [4]).
In order to add queue-dependency to the previous Hawkes approach, we proceed in much the same way as we did in the previous section except that the order book state is no longer represented by a single queue quantity q (that represented either the best ask or the best bid quantity) but by both the best bid and the best ask queue sizes (q_a, q_b). Let us point out that now the queue sizes q_a and q_b, by definition, never reach zero, moreover each process dN^ℓ encodes the full history of all the events of type ℓ happening at the corresponding best side, independently of the various moves of the best ask/bid prices. So adding queue-dependency to such a model calls for a mechanism that is able to modulate (as a function of the orderbook state) not only the exogenous intensity μ_ℓ (as in the previous model) but also the Hawkes part of the intensity. To illustrate that point, we can mention a situation where obviously the Hawkes kernel matrix should explicitely depend on the queue states: This is for instance the case of P^+ events which are very unlikely to occur (for a large tick asset) when the ask queue is big and the bid queue is small.
We consider the simplest possibility where both the exogenous and the self-exciting part of the intensity share the same multiplicative dependence on the states and introduce the following model, referred to in the following as QRH-II

$$\lambda^\ell(t) = f^\ell(q_a(t), q_b(t)) \left(\mu^\ell + \sum_m \int_0^t \phi^{\ell m}(t-s) dN_s^m \right), \tag{17}$$

where the functions f^ℓ that encode the dependence on the orderbook states, modulate not only the exogenous intensity (as in Eq. (4)) but also the Hawkes term. Let us notice that unlike QRH-I model, the QRH-II model is mainly a model for the order flow, so we disregard the relationship between the order arrivals and the queue sizes and consider these queues as (observable) exogenous processes. As before, we choose a parametric form for the kernels $\phi^{\ell m}$ and in particular we adopt the same exponential-sum specification as provided in Eq. (6).

[2] We recall that the midprice corresponds to the average of the best ask price with the best bid price.

3.2 From log-likelihood estimation to least square estimation

Once the parametric form (6) for the kernels has been specified, the model can be estimated using MLE. Although the QRH-II model slightly differs from the QRH-I model, the calculation of its log likelihood function and its gradients follows the same track presented in Appendix A.2, with only a trivial modification required. Moreover, thanks to the chosen parametrization, the log-likelihood is again a convex function of the parameters, thus guaranteeing the existence of a global optimum.
Since the number of configurations of the orderbook states grows quadratically in the number of states considered per-side, this can become problematic if this number is too large. Considering every possible state (i.e., every possible values for each queue) would require very long computation times. To limit the number of configurations, we thus consider the orderbook to be in state (q_a^i, q_b^j) if the bid queue size is within its i-th quintile (i.e. the $25 \cdot i$ percentile) and the ask queue is in its j-th quintile. So the estimated function f^ℓ is actually a function of the quintiles $f^\ell(q_a^i, q_b^j)$. Finally, we choose the normalization (note that any other normalization would be equivalent up to a rescaling of the kernels and of the μ^ℓ in Eq. (17))

$$f^\ell(q_a^1, q_b^1) = 1. \tag{18}$$

Using maximum likelihood, we calibrate the so-obtained model on the same dataset used in the previous section. As shown in Tables 4 and 5, our model outperforms in terms of goodness of fit the pure-Hawkes model introduced in [4] when calibrated with sum of exponential kernels. In particular, a likelihood ratio test rejects the null hypothesis of a pure-Hawkes model with a p-value $< 10^{-16}$.

Bund				
	L	AIC	BIC	# parameters
QRH-II	5.348×10^8	-1.070×10^9	-1.070×10^9	400
Hawkes	5.200×10^8	-1.040×10^9	-1.040×10^9	200
DAX				
	L	AIC	BIC	# parameters
QRH-II	4.626×10^8	-9.253×10^8	-9.253×10^8	400
Hawkes	4.488×10^8	-8.976×10^8	-8.976×10^8	200

Table 4: Log-likelihood, AIC, and BIC values for the QRH-II model (defined by (17)) and the Hawkes model (defined in [4]) for Bund and DAX data.

Bund			
	LR	df	p-value
H_0 = Hawkes, H_1 = QRH-II	$2.9 \cdot 10^7$	200	$< 10^{-16}$
DAX			
	LR	df	p-value
H_0 = Hawkes, H_1 = QRH-II	$2.8 \cdot 10^7$	200	$< 10^{-16}$

Table 5: Likelihood ratio test statistic and p-values for the case where the null hypothesis is the QRH-II model (defined by (17)) and for the case where the null hypothesis is the Hawkes model (defined in [4]).

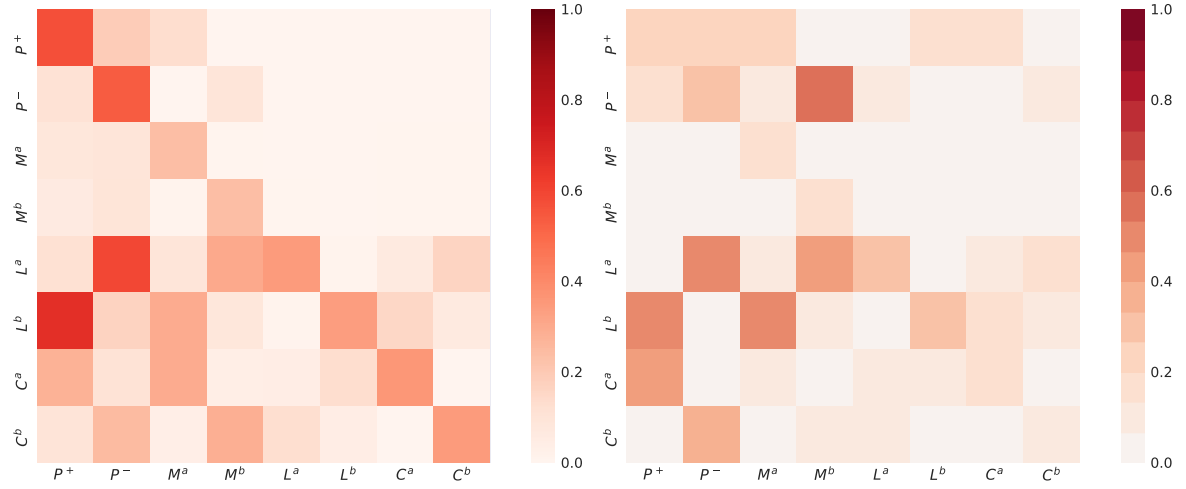

Figure 8: MLE Kernel norm matrices $\int \phi^{lm}(t)dt$ for Bund future (left) and DAX future (right) for the QRH-II model (defined by (17)).

A detailed discussion on the so-obtained MLE results will be made in Section 3.3. For now, for the sake of simplicity, let us first put apart the effect of the queue dependency and just look at the kernel estimation. Figure 8 represents the so-obtained kernel norm matrices $\{\int \phi^{\ell m}(t)dt\}_{\ell m}$ given by our model. This figure has to be compared with the matrix obtained in Figure 4 of [4] for which the same model (without queue dependency) has been used. Let us point out that in order to obtain this matrix, [4] performed a non-parametric estimation which allows negative values for the kernels and consequently negative values for some elements of the matrix $\{\int \phi^{\ell m}(t)dt\}_{\ell m}$. These negative values account for inhibition dynamics, i.e., decreasing the intensity of a given type of event. One could show that negative values for a kernel could lead at a finite time (and with a non zero probability) to some situation where the sum (17) is negative leading to a negative intensity which, of course, does not make any sense. In order to circumvent this difficulty, it is common to replace equation (17) by a non-linear equation of the form

$$\lambda^{\ell}(t) = f^{\ell}(q_a(t), q_b(t)) \left(\mu^{\ell} + \sum_m \int_0^t \phi^{\ell m}(t-s) dN_s^m \right)^+, \tag{19}$$

where the operator $(\ldots)^+$ does not change the value of the quantity in between the parenthesis in the case this quantity is positive and is 0 if this quantity is negative. However, in the framework of this non linear Hawkes model, MLE becomes intractable since it is no longer a convex problem. As explained in [14], this problem is tractable in some sense when considering the least square approach (see also [24] for example of least square estimation with negative valued kernels).
Thus, the least square estimations allows negative valued kernels whereas, in the MLE framework as presented above, we forced all the kernels (i.e., all the α_u^{lm}'s in Eq. (6)) to be positive valued. This mainly explains the differences found between Figure 4 in [4] and our Figure 8. For the sake of just naming one striking difference, in [4], the kernel integral $\int \phi^{L^b P^-}$ (resp. $\int \phi^{L^a P^+}$) is found to be strongly negative. Actually, as seen in Appendix B.6 in [4], the kernel itself $\phi^{L^b P^-}$ (resp. $\phi^{L^a P^+}$) is mostly negative at all time scales. This fact can be seen as a natural dynamic induced by market makers: when the price goes down, the efficient price is closer to the best bid price thus less limit orders are placed on the bid size (the gain is small compared to sending an aggressive order) and more limit orders are placed on the ask side.
Let us point out that forcing the kernels to have only positive values (i.e., forcing all the α_u^{lm} to be positive as assumed when performing MLE) will a priori not only lead to highly biased values for kernels with negative values but is likely also to induce high bias for kernel with only positive values since the estimation performed is a joint estimation of all the kernels involving intricate relationships between these kernels (see [6] for examples of such biases).
Consequently, it appears that one should use least square based estimation rather than MLE. Details about least square estimation can be found in Appendix A.3. It consists in minimizing $R(\theta)$ as defined

by

$$R(\theta) = \sum_{\ell=1}^{D} R^{\ell}(\theta), \quad \text{with} \quad R^{\ell}(\theta) = \int_0^T \lambda_{\ell}^2(t;\theta|\mathscr{F}_t)dt \; - \sum_{k=1}^{N^{\ell}} \lambda_{\ell}(t_k^{\ell};\theta|\mathscr{F}_{t_k^{\ell}}) \tag{20}$$

where λ_ℓ is given by Eq. 17[3]. One could easily verify that this problem is convex as a function of the $\alpha_u lm$ (see Eq. (6)), thus the existence of a global optimum is guaranteed. As shown in Appendix A.3, this parametrization allows a computationally efficient calculation of the squares loss function R together with its gradient.

3.3 Fitting results and comments

In this section we present and comment the results obtained through least square estimation of the QRH-II model as defined by Eq. (17).

Estimation of the $f^\ell(q_a, q_b)$ functions. In our results we document a clear dependence of the order arrival rates on both q_a^i and q_b^j, indicating that the state of the LOB has a clear influence on the order arrival rates. Let us point out that previous works (e.g., [23] and [18]) suggest that this dependence is actually a dependence on the imbalance of the queue sizes as defined by

$$I(t) = \frac{v_b(t) - v_a(t)}{v_b(t) + v_a(t)} = \frac{q_b(t) - q_a(t)}{q_b(t) + q_a(t)} \tag{21}$$

where $v_{a/b}(t)$ denote the volume available at time t at best ask/bid prices and we have assumed that orders have a constant volume corresponding to the AES as defined previously. The imbalance represents the simplest proxy to account for the instantaneous buying pressure. In that respect, in Figures 9 and 10 we plot the parameters $f^\ell(q_a^i, q_b^j)$, in logarithmic scale, for each order type ℓ as a function of the imbalance I calculated as the median imbalance in the state interval associated with quantiles (q_a^i, q_b^j). Let us note that in these plots, the dot sizes indicate the sizes of the corresponding q_b^i. By looking at Figures 9 and 10 we can make the following observations: First, the variation of the imbalance on the order book captures most of the variations of the intensive parameters $f^\ell(q_a, q_b)$, this is more evident on the large tick asset (Bund) for which f can span almost three order of magnitudes (for P, and M events) as I ranges from -1 to $+1$. The variations of $f^\ell(q_a, q_b)$ are smaller for the small tick asset (DAX) so the effect of the imbalance is less pronounced albeit still visible for P and to a lesser extent L events. This is in line with the observation that the imbalance is a very good predictor for midprice changes for large tick assets while its predictive power is less marked for small tick ones (see e.g. [12]). One aspect to keep in mind while analyzing these figures is that, especially for a small tick asset, there is also a considerable amount of information in the deeper levels of the book that is not taken into account here.
In Figure 9 corresponding to the Bund results, we observe that for midprice changes (P events) there seems to be a kind of threshold effect, in that for imbalance values below $I = -1/2$, upwards price increments are dramatically inhibited, while large positive imbalance only marginally increases the likelihood of upwards price changes. This is rather natural since an imbalance smaller than $-1/2$ corresponds to the case the ask size is at least three times bigger than the bid size which makes upwards move of the midprice very unlikely. More surprinsingly, the agents seem to strongly condition their decision to use aggressive orders (M events) on the state of the order book. Maybe they rush to get last available liquidity before an upwards midprice move (indeed f gets very large for M events when the imbalance is large, i.e., when there is hardly anything left on the ask size). Limit and Cancel orders appear to be much less sensitive to the state of the book. For the DAX, as we already pointed out, the dependence on the imbalance is much less pronounced (see Figure 10) and, expcept the mid-price changes, all factor $f^\ell(q_a, q_b)$ are weaky varying with I. Notice hower that for market (M) and cancel (C) orders we observe a regime switching around $I = 1/2$. These regimes correspond respectively to quite large and very small ask queue size. In the latter case (that turns out to occur when the spread equals one tick) the occurence of market and cancel orders at ask that do not change the midprice is very unlikely.
Last but not least, let us remark that the intensity variations that are not captured by the imbalance are mainly located around $I = 0$ where the individual queue sizes (q_a or q_b which are almost equal) seem to have an important impact. Thus, for instance, for the Bund, $f^{P^+}(q, q)$ seems to increase with q. Actually, this can be explained by a more thorough analysis that shows that the average spread is increasing with q and is very close to 2 when q is large. The same kind of remarks could be done for market orders.

[3] or alternatively by Eq (19) in the sense given in [14]

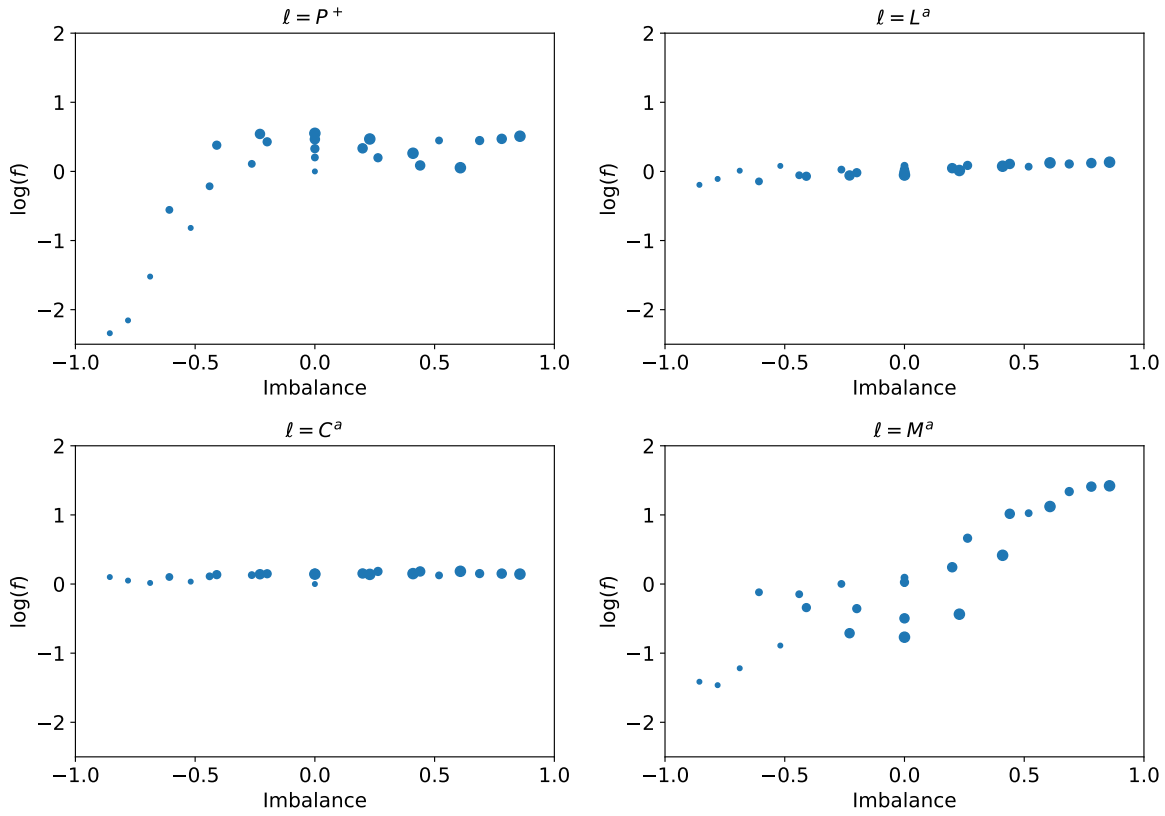

Figure 9: From left to right, upside to downside, $\log_{10}(f^l(q_a^i, q_b^j))$ for $l = P^+, L^a, C^a, M^a$ of QRH model as a function of the imbalance 21, Bund future. The quantiles are the same for bid and ask sides and correspond to $q_a^1 = q_b^1 =]0, 80]$, $q_a^2 = q_b^2 =]80, 165]$, $q_a^3 = q_b^3 =]165, 258]$, $q_a^4 = q_b^4 =]258, 386]$ and $q_a^5 = q_b^5 =]386, +\infty[$.

As this example illustrates other microstructural variables and notably the spread, should be taken into account for an even more complete model. This however is outside the scope of the present work.

Comparison of estimated and empirical intensity. To further validate the QRH-II model, we test its ability to reproduce the empirical intensity. The MLE of the averaged intensity conditioned by the state $q = (q_a, q_b)$ 15 writes (in a non parametric framework)

$$\hat{\Lambda}^\ell(q) = \hat{\Lambda}(q) \frac{\sum_{t_k^\ell} \mathbb{1}_{q(t_k^{l-})=q}}{\sum_{t_k} \mathbb{1}_{q(t_k^-)=q}} \quad \text{with} \quad \hat{\Lambda}(q) = \text{mean}(t_k - t_{k-1} | q(t_k^-) = q)^{-1}, \tag{22}$$

where the operator mean(...) corresponds to the empirical mean. This estimation is simply based on the observations of the process $\{N_t^l\}_l$. One can compute a corresponding estimator $\hat{\Lambda}^\ell_{\text{QRH-II}}(q)$ using QRH-II parametric form of $\hat{\lambda}^\ell_{\text{QRH-II}}(t_k^{\ell,-}|q)$, this leads to

$$\hat{\Lambda}^\ell_{\text{QRH}}(q) = \text{mean}(\hat{\lambda}^\ell_{\text{QRH-II}}(t_k^{\ell,-}) | q(t_k^{\ell,-}) = q) \tag{23}$$

In order to synthesize the so-obtained results, we choose not to present the comparison of $\hat{\Lambda}^\ell_{\text{QRH-II}}(q)$ with $\hat{\Lambda}^\ell(q)$ for all types of orders and for all states q. Instead, for each type of order, we report the weighted relative error Δ^ℓ defined as:

$$\Delta^\ell = \frac{\sum_q \left| \hat{\Lambda}^\ell(q) - \hat{\Lambda}^\ell_{\text{QRH-II}}(q) \right| N^\ell(q)}{\sum_q \hat{\Lambda}^\ell(q) N^\ell(q)} \tag{24}$$

The weighted relative error for Bund and DAX is presented in table 6. We observe that Δ^ℓ are of the order of 10%, which provides a satisfactory match to the empirically observed intensity.

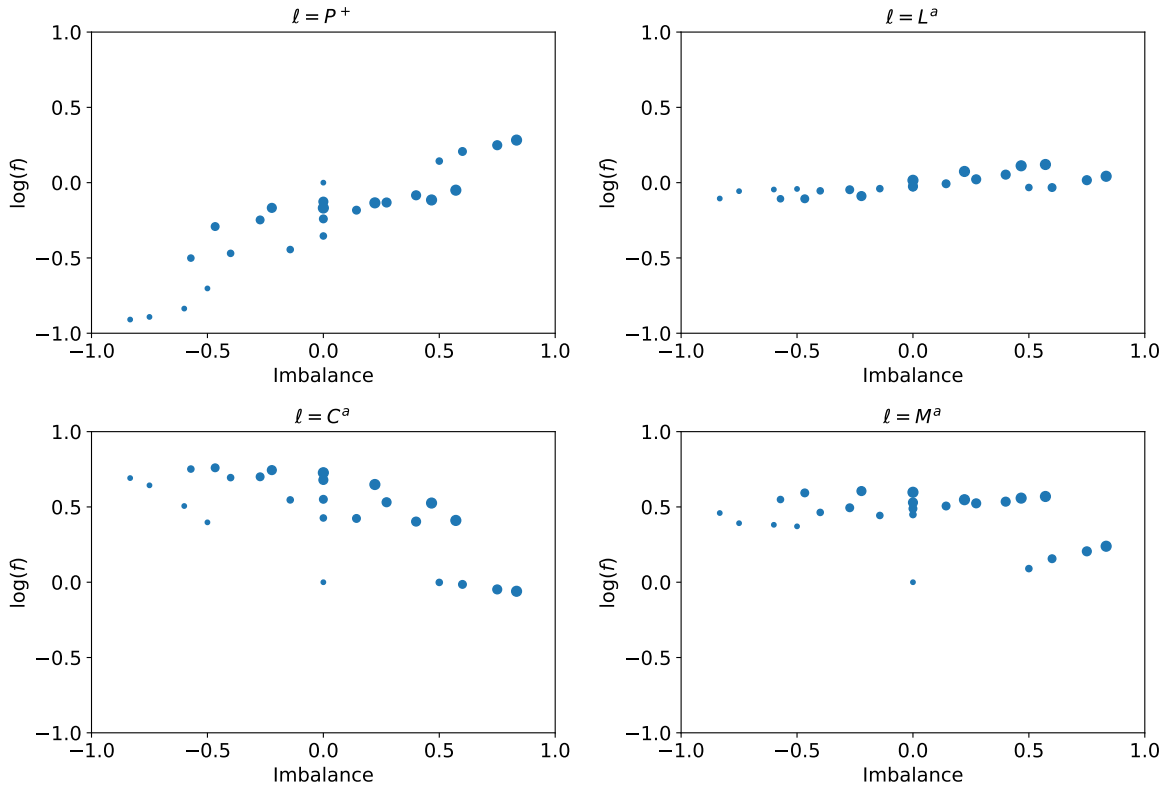

Figure 10: From left to right, upside to downside, $\log_{10}(f^l(q_a^i, q_b^j))$ for $l = P^+, L^a, C^a, M^a$ of QRH model as a function of the imbalance 21, DAX index future. The quantiles are the same for bid and ask sides and correspond to $q_a^1 = q_b^1 =]0, 2]$, $q_a^2 = q_b^2 =]2, 3]$, $q_a^3 = q_b^3 =]3, 5]$, $q_a^4 = q_b^4 =]5, 8]$ and $q_a^5 = q_b^5 = [8, +\infty[$.

	P^+	P^-	L^a	L^b	C^a	C^b	M^a	M^b
Bund	14.2%	10.5%	6.7%	6.0%	7.4%	8.1%	4.0%	12.0%
DAX	8.2%	5.9%	0.5%	4.7%	7.1%	1.1%	1.6%	5.9%

Table 6: Error of average intensities by order type

Analysis of the kernel norm matrices. Finally, to complete the analysis of our results, in Figure 11 we display the matrices of the estimated norms $\{\int \phi^{\ell m}(t)dt\}_{\ell m}$ for both the Bund and the DAX. These matrices provide information on the average interactions between different event types when queue dependence is disregarded. As such they are the counterpart of the kernel norm matrices shown in Figure 4 of [4]. We recover a lot of the features highlighted in [4], such as the strong diagonal components for limit, market and cancel orders (a signature of order splitting), as well as the fact that market orders and price movement appear to influence much more limit and cancel than the other way round. Notably, for the Bund, when the midprice moves up (resp. down) it decreases (resp. increases) the rate of limit orders on the ask (resp. best) side since the efficient price is close to the best ask, there is no gain to send a limit versus a market. This also explains why it increases (resp. decreases) the rate of cancel orders sent on the best ask (resp. bid) side. The same effect can be more or less seen (but attenuated a lot) on the DAX. It is attenuated because the tick is small on the DAX, so the efficient price argument is not as strong. Let us point out that the exact same argument can be used to explain the effects of market orders on limit and cancel orders. The influence of market order over price change is mainly because that under our settings, market order consumes the liquidity at the best prices, which could eventually create a new price. Since DAX is a small tick asset and queue size at the best prices is smaller that Bund, market orders are more likely to generate new prices. So the influence of market order over price change is more visible. We can see that for DAX future, limit orders and cancellations also has stronger influence toward price changes, for the same reason. One can also see that, apart from the self-exciting effect due to splitting of orders, for the Bund, limit orders on one side are coupled with cancel orders on the other side. This can be seen as a simple market maker strategy (rebalancing of the position). It

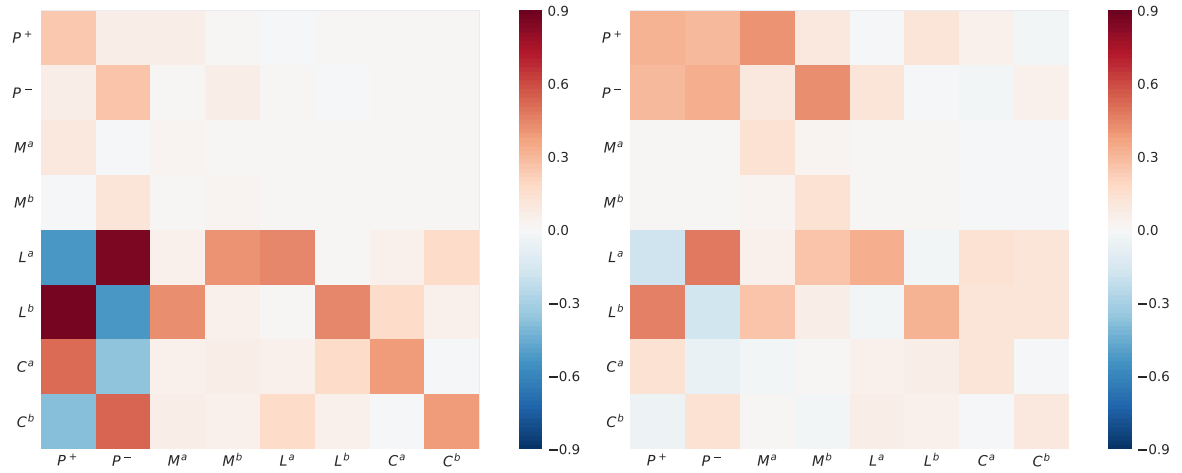

Figure 11: The estimated matrix norms $\int \phi^{lm}(t)dt$ using least square estimation QRH-II model. Bund future on the left and DAX future on the right.

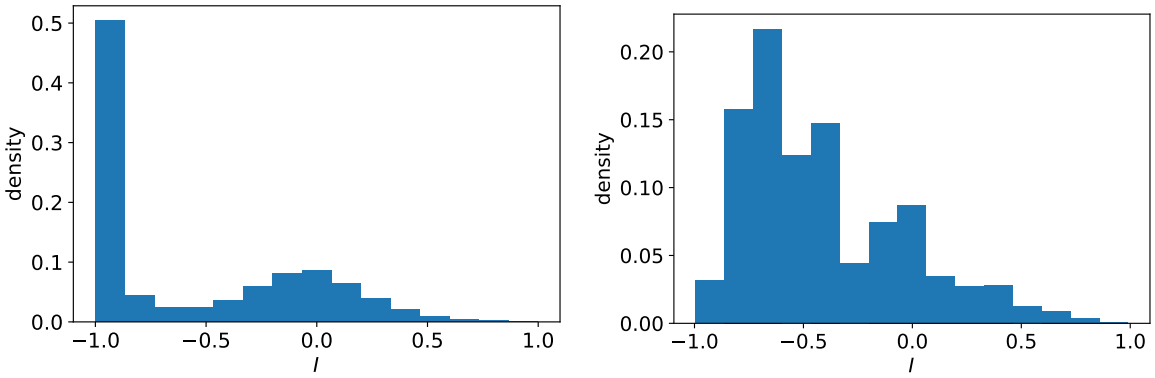

Figure 12: Empirical frequencies of observed imbalance values right after a P^+ event for the Bund (left panel) and the DAX (right panel). Notice the large components for negative imbalance values in both cases.

does not show on the DAX certainly because, since the tick size is much smaller, the same rebalancing strategy does not affect necessary both best sides. We refer the reader to [4], Section 4 for further details on the interpretation of these features.

	Rate of mean-reversion
Bund	0.65
DAX	0.56

Table 7: Measure of mean-reversion of price: Empirical probabilities than two successive midprice change events have opposite directions.

Though for most features, QRH-II model (Figure 11) and the pure Hawkes model (Figure 4 in [4])) are similar, we can however observe a striking difference in the $P \to P$ and $P \to T$ submatrices. We can notice in QRH-II the absence of a strong excitation between P^+ to P^- and P^- to P^+ (top left 2×2 submatrices in Figure 11) which should be the signature of the high frequency price mean reversion (as explained in [4]). For a pure Hawkes processes model, like in work [4], the mean-reversion of price is reflected on the strong anti-diagonal terms of the $P \to P$ kernel norms submatrix (i.e., strong $P^+ \to P^-$ and $P^- \to P^+$ terms). This results from the fact that (in average) a P^+ event will generate more P^- events than P^+ events, and vice versa. As shown in table 7, the actual midprice series are strongly mean-reverting, so it is likely that, within the QRH-II model this feature should be explained by the queue-reactive function $f(q_a, q_b)$. Indeed, as illustrated in Fig. 12, after a midprice change the imbalance evolves in favor of a price move towards the opposite direction. For instance, an upward price jump

P^+ leads, most of the time, to a negative imbalance either because of a refill of the best ask queue q_a or simply because a single buy limit order is sent within the spread. According to results of Figs. 9 and 10, in this case we have $f^{P^+}(q_a, q_b) \ll 1$ and $f^{P^-}(q_a, q_b) > 1$. Conversely, after a downward price change, P^-, we will have $f^{P^-}(q_a, q_b) \ll 1$ and $f^{P^+}(q_a, q_b) > 1$. The 2×2 submatrix $P \to P$ estimated within a pure Hawkes model is then likely to correspond to the QRH-II Hawkes submatrix multiplied by a large factor on its anti-diagonal and a small factor on its diagonal. The same kind of argument based on imbalance impact can be invoked to explain the high diagonal values of the $P \to T$ submatrix while the highest values in [4] were rather observed on the anti-diagonal.

4 Summary and prospects

In this paper, we introduced two "Queue Reactive Hawkes models" (QRH-I and QRH-II) with the ambition to improve respectively the approach of Huang et al. [16] on the queue reactive nature of the LOB dynamics and the model of Bacry et al. [4]. We show that shuch models can be easily calibrated within a parametric approach. Our empirical findings on two different future asset from Eurex, namely Bund and DAX order book data, suggest that both queue reactive and past order flow dependencies are relevant to account for the occurrence likelihood of future order book events. Indeed, both models outperforms in terms of goodness of fit a pure Hawkes model as well as a pure queue-reactive one. As far as QRH-I model is concerned, our framework allows one to remain within the framework of Markov processes that has ergodic properties so, along the same line as in Huang et al. approach, we can define and estimate the queue size distribution associated with the invariant measure of the model. The QRH-II model lead us to refine Bacry et al. findings [4] by accounting for the states of best bid and best ask queues. It turns out that the volume imbalance allows us to explain most of the queue dependence.
The QRH-II model could be improved by accouting for the interactions between the queue sizes (and thus the imbalance) and the order flow, as in the QRH-I model and also to include an explicit dependence on the spread for small tick assets. Some substantial simplifications we made could also be removed in order to have a even more realistic model such as, for instance, dropping the assumption of unitary order sizes by adopting a similar approach to [19]. Besides considering various applications of these models to design and optimize high frequency trading strategies, from a numerical point of view, it remains to define approaches that allow one handle the full QR-model as defined in Eq. (2) where not only the queue dependencies of the exogenous and Hawkes kernels are arbitrary but that simultaneously account for all the order book queues up to a given depth. Maybe some recent approaches and techniques developed in statistical learning would be helpful to handle such high dimensional problems. From a mathematical point of view, it remains to develop a deeper understanding of the stability and stationarity conditions for queue dependent Hawkes models. More fundamentally, a clear understanding to the observed shapes of the exogenous intensities and the imbalance dependence of the order flow arrival rates in term of the (rational) behavior of various market participants remain open questions.

A Appendix

A.1 Proof of the existence of invariant distribution

We are motivated to prove the existence of invariant distribution of QRH-I model presented in Part I, for the reason that if such invariant distribution exists, we could approximate the empirical distribution of queue size by simulating the QRH-I model for a long time. The proof is made with the help of Lyapunov function. First let's define:

$$o_{\ell mu}(t) = \int_0^t \alpha_u^{\ell m} e^{-\beta_u (t-s)} dN_s^m \tag{25}$$

By adopting defintion (23) and (6), the intensity function now takes the form:

$$\lambda^\ell(t) = \mu^\ell(q(t^-)) + \sum_{m=1}^{D} \sum_{u=1}^{U} o_{\ell mu}(t) \ , \tag{26}$$

up to an overall factor $\mathbb{1}_{q(t)\neq 0}$ that has to be considered for orders that consume the queue size. We note J the set of order types that increase the queue size and I the set of order types that decrease the queue

size. We further assume that $\sum_{\ell\in I}\mu^\ell(q)\geq c_\ell q$ and $\mu^m(q)\leq c^*, \forall m\in J$. The queue size is determined by the sum of order flows:

$$q(t)=\sum_{m\in J}N_m-\sum_{\ell\in I}N_m \tag{27}$$

Let $\vec{o}(t)$ be the vector obtained as a vertical stacking of the components $o_{\ell mu}(t)$ for all $(\ell,m,u)\in\{1,\dots,D\}^2\times\{1,\dots U\}$. It is not difficult to verify that the vector process $(q,\vec{o})^T$ is Markovian. We then aim at constructing a Lyapunov function for $(q,\vec{o})^T$. To start with, we construct Lyapunov functions for q and $\vec{o}$ seperately, then we combine them together.

A.1.1 Lyapunov function for $\vec{o}$

Let us first write the differential form of $\vec{o}(t)$:

$$\mathrm{d}o_{\ell mu}(t)=-\beta_u o_{\ell mu}(t)\mathrm{d}t+\alpha_u^{\ell m}\mathrm{d}N_t^m \tag{28}$$

Then for any arbitrary suitable function F who maps $\mathbb{R}^{2D+U}$ to $\mathbb{R}$, the infinitesimal generator takes the form:

$$\mathcal{L}F(\vec{o})=\sum_m\lambda_m(F(\vec{o}+\Delta_m(\vec{o}))-F(\vec{o}))-\sum_{\ell,m,u}\beta_u o_{\ell mu}\frac{\partial F}{\partial o_{\ell mu}} \tag{29}$$

Where λ_m is the probability for an event to arrive in dimension m of the point process $\vec{N}$, and $\Delta_m(\vec{o})$ is jump of $\vec{o}$ caused by this event. We then define the matrix A as:

$$A_{\ell m}=\sum_u\frac{\alpha_u^{\ell m}}{\beta_u} \tag{30}$$

We assume that the dynamics of order flows described by the Hawkes part of the QRH-I model corresponds to a stable Hawkes process. According to Perron-Frobenius theorem, the maximal eigenvalue κ of A satisfies $0<\kappa<1$, and $\vec{\epsilon}$ the associated eigenvector of κ satisfies $\forall l,\epsilon_l>0$. We then note

$$\delta_{\ell mu}:=\delta_{\ell u}=\frac{\epsilon_\ell}{\beta_u} \tag{31}$$

We choose function V_1 of $\vec{o}$ as:

$$V_1(\vec{o})=\sum_{\ell,m,u}\delta_{\ell mu}o_{\ell mu} \tag{32}$$

With notations defined above, we could then verify that

$$\begin{aligned}
\mathcal{L}V_1(\vec{o}) &= \sum_m(\mu_m+\sum_{p,q}o_{mpq})\mathbb{1}_{\ell=0\vee q(t)>0}\sum_{\ell,u}\delta_{\ell mu}\alpha_u^{\ell m}-\sum_{\ell,m,u}\beta_u o_{\ell mu}\delta_{\ell mu}\\
&\leq \sum_m(\mu_m+\sum_{p,q}o_{mpq})\sum_{\ell,u}\delta_{\ell mu}\alpha_u^{\ell m}-\sum_{\ell,m,u}\beta_u o_{\ell mu}\delta_{\ell mu}\\
&= \sum_{\ell,m,u}\mu_m\delta_{\ell mu}\alpha_u^{\ell m}+\sum_m(\sum_{p,q}o_{mpq})(\sum_{\ell,u}\frac{\epsilon_\ell}{\beta_u}\alpha_u^{\ell m})-\sum_{\ell,m,u}\beta_u o_{\ell mu}\delta_{\ell mu}\\
&= C_1+\sum_m(\sum_{p,q}o_{mpq})(\sum_\ell\epsilon_\ell\sum_u\frac{\alpha_u^{\ell m}}{\beta_u})-\sum_{\ell,m,u}\beta_u o_{\ell mu}\delta_{\ell mu}\\
&= C_1+\sum_m(\sum_{p,q}o_{mpq})\epsilon_m\kappa-\sum_{\ell,m,u}\epsilon_\ell o_{\ell mu}\\
&= C_1-(1-\kappa)\sum_{\ell,m,u}\epsilon_\ell o_{\ell mu}\\
&= C_1-(1-\kappa)\sum_{\ell,m,u}\beta_u\delta_{\ell mu}o_{\ell mu}\\
&\leq -\rho_1 V+C_1
\end{aligned} \tag{33}$$

Where the constant ρ_1 is chosen as

$$\rho_1=(1-\kappa)\inf\beta_u \tag{34}$$

A.1.2 Lyapunov function for q

Similarly, let us first write the differential form of $q(t)$:

$$\mathrm{d}q(t) = \sum_{m\in J} \mathrm{d}N_t^m - \sum_{\ell\in I} \mathrm{d}N_t^\ell \tag{35}$$

We keep using the same notation of Δ as in Eq 29. For any arbitrary suitable function F who maps $\mathbb{R}^+$ to $\mathbb{R}$, the infinitesimal generator takes the form:

$$\mathcal{L}F(q) = \sum_m \lambda_m(F(q+\Delta_m(q)) - F(q)) \tag{36}$$

Next, we choose V_2 as the identity function of q:

$$V_2(q) := q \tag{37}$$

Then it could be easily verified that $V_2(q)$ is a Lyapunov function for q:

$$\begin{aligned}
\mathcal{L}V_2(q) &= \sum_{m\in J} \lambda_m(q) - \sum_{\ell\in I} \lambda_\ell(q) \\
&\le \sum_{m\in J} \mu_m(q) - \sum_{\ell\in I} \mu_\ell(q) + \sum_{\ell,m,u} o_{\ell mu} \\
&\le \sum_{m\in J} c^* - \sum_{\ell\in I} c_\ell q + \sum_{\ell,m,u} o_{\ell mu} \\
&\le -\sum_{\ell\in I} c_\ell q + C_2 \\
&\le -\rho_2 V_2 + C_2
\end{aligned} \tag{38}$$

Remind that constants c^*, c^ℓ are defined before. The constant ρ_2 is simply chosen as $\rho_2 = \max_\ell c_\ell$. Also, it must be noted that here

$$C_2 := \sum_{m\in J} c^* + \sum_{\ell,m,u} o_{\ell mu} \tag{39}$$

which depends on $\vec{o}$. Another remark is since m and c^* are fixed, there must exist ρ^* who satisfy

$$C_2 < \rho^* \sum_{\ell,m,u} o_{\ell mu} \tag{40}$$

A.1.3 Lyapunov function for $(q, \vec{o})$

The final step is to build a function V for both q and $\vec{o}$:

$$V(q,\vec{o}) = V_2(q) + \frac{1}{\eta} V_1(\vec{o}) \tag{41}$$

We then apply previous conclusions made for V_1 and V_2. Direct calculation shows:

$$\begin{aligned}
\mathcal{L}V(q,\vec{o}) &= \mathcal{L}V_2(q) + \frac{1}{\eta}\mathcal{L}V_1(\vec{o}) \\
&= -\rho_2 V_2 + C_2 - \frac{\rho_1}{\eta} V_1 + \frac{C_1}{\eta} \\
&\le -\rho_2 V_2 + \rho^* \sum_{\ell,m,u} o_{\ell mu} - \frac{\rho_1}{\eta} V_1 + C'
\end{aligned} \tag{42}$$

Notice that since both the coefficient $\delta_{\ell mu}$ in V_1 and $o_{\ell mu}$ are positive, there must exist an η who satisfies

$$\frac{\rho_1}{2\eta} V_1(\vec{o}) > \rho^* \sum_{\ell,m,u} o_{\ell mu} \tag{43}$$

By substituting this inequality into Eq 42, we finally prove that V is a Lyapunov function for $(q, \vec{o})$:

$$\begin{aligned} \mathcal{L}V(q,\vec{o}) &\leq -\rho_2 V_2 - \frac{\rho_1}{2\eta} V_1 + C' \\ &\leq -\min(\rho_2, \frac{\rho_1}{2\eta}) V + C \end{aligned} \tag{44}$$

Given the existence of the Lyapunov function and the geometric drift condition above, under the assumption that the spectral radius of A is smaller than one, Theorem 7.1 in [20] guarantees that the process q is ergodic. Also, it converges exponentially fast towards its unique stationary distribution.

A.2 Calculation of the log-likelihood function of QRH model and its gradient

For QRH model, the log-likelihood is a function of μ and α. Let's note t_k^ℓ the timestamp of the kth event of type ℓ, and N^ℓ the total number of event of type ℓ. With such notation,

$$\begin{aligned} L(\vec{\alpha}, \vec{\mu}) = &\sum_{\ell=1}^{D} \sum_{k=1}^{N^l} \log \left(\mu^\ell(q(t_k^{\ell,-}) + \sum_{m=1}^{D} \sum_{u=1}^{U} \alpha_u^{\ell m} \beta_u \int_0^{t_k} e^{-\beta_u (t-s)} dN_s^m \right) \\ &- \sum_{\ell=1}^{D} \int_0^T \left(\mu^\ell(q(t)) + \sum_{m=1}^{D} \sum_{u=1}^{U} \alpha_u^{\ell m} \beta_u \int_0^s e^{-\beta_u (s-v)} dN_v^m \right) ds \end{aligned} \tag{45}$$

We first define g and G, whose value doesn't depend on the parameters $\vec{\mu}$ and $\vec{\alpha}$. The interest is that g and G only need to be calculated once, then they could be reused to accelerate the calculation of log-likelihood function and its gradient.

$$g_u^m(t) = \sum_{t_k^m < t} \beta_u e^{-\beta_u (t - t_k^m)} \tag{46}$$

And

$$G_u^m(t) = \int_0^t g_u^m(s) ds = \int_0^t \int_0^s \beta_u e^{-\beta_u (s-v)} dN_v^m ds \tag{47}$$

Both g and G could be calculated using the following recurrence relation:

Computation of g

$$g_u^m(t_k) = \sum_{t_{k'}^m < t_k} \beta_u e^{-\beta_u (t_k - t_{k'}^m)} \tag{48}$$

$$= \sum_{t_{k'}^m < t_{k-1}} \beta_u e^{-\beta_u (t_{k-1} - t_{k'}^m)} e^{-\beta_u (t_k - t_{k-1})} + \sum_{t_{k-1} \leq t_{k'}^m < t_k} \beta_u e^{-\beta_u (t_k - t_{k'}^m)} \tag{49}$$

$$= e^{-\beta_u (t_k - t_{k-1})} g^m(t_{k-1}) + \beta_u e^{-\beta_u (t_k - t_{k-1})} \mathbb{1}_{type(t_{k-1}^+) = m} \tag{50}$$

Computation of G

$$G_u^m(t_k) - G_u^m(t_{k-1}) = \int_{t_{k-1}}^{t_k} g_u^m(s) ds \tag{51}$$

$$= \frac{1 - e^{-\beta_u (t_k - t_{k-1})}}{\beta_u} g_u^m(t_{k-1}) + \left(1 - e^{-\beta_u (t_k - t_{k-1})}\right) \mathbb{1}_{type(t_{k-1}^+) = m} \tag{52}$$

We will see later that we only need the value of g and G at for every t_k.

A.2.1 Log-likelihood function

By exploiting the definition of g and G, the log-likelihood function could be rewritten in the following way:

$$\begin{aligned} L(\alpha,\mu) = &\sum_{\ell=1}^{D}\sum_{k=1}^{N^\ell}\log\Big(\mu^\ell(q(t_k^{\ell,-})) + \sum_{m=1}^{D}\sum_{u=1}^{U}\alpha_u^{\ell m} g_u^m(t_k)\Big) \\ &- \sum_{\ell=1}^{D}\sum_{k=1}^{N}\mu^\ell(q(t_k^-))(t_k - t_{k-1}) - \sum_{\ell=1}^{D}\mu^\ell_{q(t_N^+)}(T - t_N) \\ &- \sum_{\ell=1}^{D}\sum_{k=1}^{N}\sum_{m=1}^{D}\sum_{u=1}^{U}\alpha_u^{\ell m}(G_u^m(t_k) - G_u^m(t_{k-1})) - \sum_{\ell=1}^{D}\sum_{m=1}^{D}\sum_{u=1}^{U}\alpha_u^{\ell m}(G_u^m(T) - G_u^m(t_N)) \end{aligned} \tag{53}$$

A.2.2 Gradients

With a little abuse use of the symbol q and $q(t)$, direct calculation shows that:

$$\frac{\partial L}{\partial \mu^\ell(q)} = \sum_{k=1}^{N^\ell}\frac{\mathbb{1}_{q(t_k^{\ell,-})=q}}{\mu^\ell(q(t_k^{\ell,-})) + \sum_{m=1}^{D}\sum_{u=1}^{U}\alpha_u^{\ell m} g_u^m(t_k^\ell)} - \sum_{k=1}^{N}\mathbb{1}_{q(t_k^-)=q}(t_k - t_{k-1}) - \mathbb{1}_{q(t_N^+)=q}(T - t_N) \tag{54}$$

$$\frac{\partial L}{\partial \alpha_u^{\ell m}} = \sum_{k=1}^{N^\ell}\frac{g_u^m(t_k^\ell)}{\mu^\ell(q(t_k^{\ell,-})) + \sum_{m=1}^{D}\sum_{u=1}^{U}\alpha_u^{\ell m} g_u^m(t_k^\ell)} - \sum_{k=1}^{N}(G_u^m(t_k) - G_u^m(t_{k-1})) - (G_u^m(T) - G_u^m(t_N)) \tag{55}$$

A.3 Calculating the least squares function of QRH model and its gradient

A.3.1 Least squares function

For the QRH model described in the Part II. Let's note q the state of the LOB by combining the state of both the ask side and the bide side.

$$q = q_a \times q_b \quad f(q(t)) := f(q_a(t^-), q_b(t^-)) \tag{56}$$

Then at time t, the intensity function of dimension ℓ is:

$$\lambda_t^\ell = f^\ell(q(t))\Big(\mu^\ell + \sum_m\sum_u \alpha_u^{\ell m}\int_0^t \beta_u e^{-\beta_u(t-s)} dN_s^m\Big) \tag{57}$$

Similar to the calculation of log-likelihood funtion and its gradients, we first define some intermediate vairables whose value doesn't depend μ, α or f. They only need to be calculated once in the pre-processing stage. Then they could be reused during the calculation of least squares function and its gradient.

$$g_u^\ell(t) = \sum_{t_k^\ell < t}\beta_u e^{-\beta_u(t - t_k^\ell)} \tag{58}$$

$$G_u^m(q) = \int_0^T g_u^m(s)\mathbb{1}_{type(q(t))=q} dt \tag{59}$$

$$H_{uu'}^{mm'}(q) = \int_0^T \mathbb{1}_{type(q(t))=q} g_u^m(t) g_{u'}^{m\prime}(t) dt \tag{60}$$

$$D(q) = \int_0^T \mathbb{1}_{type(q(t))=q} dt \tag{61}$$

$$C^m(q) = \sum_{k=1}^{N^m}\mathbb{1}_{type(t_k^{m,-})=q} \tag{62}$$

Computation of g

$$g_u^m(t_k) = \sum_{t_{k'}^m < t_k} \beta_u e^{-\beta_u(t_k - t_{k'}^m)} \tag{63}$$

$$= \sum_{t_{k'}^m < t_{k-1}} \beta_u e^{-\beta_u(t_{k-1} - t_{k'}^m)} e^{-\beta_u(t_k - t_{k-1})} + \sum_{t_{k-1} \le t_{k'}^m < t_k} \beta_u e^{-\beta_u(t_k - t_{k'}^m)} \tag{64}$$

$$= e^{-\beta_u(t_k - t_{k-1})} g^m(t_{k-1}) + \beta_u e^{-\beta_u(t_k - t_{k-1})} \mathbb{1}_{type(t_k^-)=m} \tag{65}$$

Computation of G

$$G_u^m(t_k) - G_u^m(t_{k-1}) = \int_{t_{k-1}}^{t_k} g_u^m(s) ds \tag{66}$$

$$= \frac{1 - e^{-\beta_u(t_k - t_{k-1})}}{\beta_u} g_u^m(t_{k-1}) + \left(1 - e^{-\beta_u(t_k - t_{k-1})}\right) \mathbb{1}_{type(t_{k-1}^+)=m} \tag{67}$$

Computation of H

$$H_{uu'}^{mm'}(q) = \sum_k \int_{t_{k-1}}^{t_k} \mathbb{1}_{type(q(t))=q} g_u^m(t) g_{u'}^{m\prime}(t) dt \tag{68}$$

$$= \sum_k \mathbb{1}_{type(q(t_k^-))=q} g_u^m(t_{k-1}) g_{u'}^{m\prime}(t_{k-1}) e^{-(\beta_u + \beta_{u'})(t_k - t_{k-1})} \tag{69}$$

A.3.2 Least squares function

With the intermidiate variables defined in the previous part, we can rewrite the least squares function defined in 20 in a more efficient form for calculation. In dimension ℓ,

$$\begin{aligned} R^\ell(\theta) &= \int_0^T \lambda_\ell^2(t;\theta|\mathscr{F}_t) dt \; - \sum_{k=1}^{N^\ell} \lambda_\ell(t_k^\ell;\theta|\mathscr{F}_{t_k^\ell}) \\ &= \int_0^T f^\ell(q(t))^2 \Big(\mu^\ell + \sum_m \sum_u \alpha_u^{\ell m} \int_0^t \beta_u e^{-\beta_u(t-s)} dN_s^m\Big)^2 dt \\ &- \sum_{k=1}^{N^\ell} f^\ell(q(t_k^\ell)) \Big(\mu^\ell + \sum_m \sum_u \alpha_u^{\ell m} \int_0^{t_k^\ell} \beta_u e^{-\beta_u(t_k^m - s)} dN_s^m\Big) \end{aligned} \tag{70}$$

Let's name the first item and the second item in the formula above *term I* and *term II* seperately. We could further decomposite *term I* into new items *term I.1*, *term I.2* and etc:

$$\begin{aligned} term\ I &= \int_0^T f^\ell(q(t))^2 \Big(\mu^\ell + \sum_m \alpha_u^{\ell m} \sum_u \int_0^t \beta_u e^{-\beta_u(t-s)} dN_s^m\Big)^2 dt \\ &= \int_0^T f^\ell(q(t))^2 {\mu^\ell}^2 dt \\ &+ \int_0^T 2 f^\ell(q(t))^2 \mu^\ell \Big(\sum_m \sum_u \alpha_u^{\ell m} \int_0^t \beta_u e^{-\beta_u(t-s)} dN_s^m\Big) dt \\ &+ \int_0^T 2 f^\ell(q(t))^2 \Big(\sum_m \sum_u \alpha_u^{\ell m} u \int_0^t \beta_u e^{-\beta_u(t-s)} dN_s^m\Big)^2 dt \end{aligned} \tag{71}$$

For *term I.1*, it could be calculated from the intermediate variables defined above:

$$\int_0^T f^\ell(q(t))^2 {\mu^\ell}^2 dt = {\mu^\ell}^2 \sum_q D(q) f^\ell(q)^2 \tag{72}$$

For *term I.2*,

$$\begin{aligned}
&\int_0^T 2f^\ell(q(t))^2\mu^\ell\Big(\sum_m\sum_u \alpha_u^{\ell m}\int_0^t \beta_u e^{-\beta_u(t-s)}dN_s^m\Big)dt \\
&=\int_0^T 2f^\ell(q(t))^2\mu^\ell\Big(\sum_m\sum_u \alpha_u^{\ell m} g_u^m(t)\Big)dt \\
&=2\mu^\ell\sum_q f^\ell(q)^2\Big(\sum_m\sum_u \alpha_u^{\ell m}\sum_u G_u^m(q)\Big)
\end{aligned} \tag{73}$$

And for *term I.3*,

$$\begin{aligned}
&\int_0^T 2f^\ell(q(t))^2\Big(\sum_m\sum_u \alpha_u^{\ell m}\int_0^t \beta_u e^{-\beta_u(t-s)}dN_s^m\Big)^2dt \\
&=\int_0^T 2f^\ell(q(t))^2\Big(\sum_m\sum_{m'}\sum_u\sum_{u'} \alpha_u^{\ell m}\alpha_{u'}^{\ell m'} g_u^m(t)g_{u'}^{m'}(t)\Big)dt \\
&=\sum_m\sum_{m'}\sum_u\sum_{u'}\alpha_u^{\ell m}\alpha_{u'}^{\ell m'}\Big(\sum_q f^\ell(q)^2 H_{uu'}^{mm'}(q)\Big)
\end{aligned} \tag{74}$$

For the convinience of notation, we flip the sign of *term II*. Then we decomposite it into *term II.1* and *term II.2*:

$$\begin{aligned}
II&=\sum_{k=1}^{N^\ell} f^\ell(q(t_k^\ell))\Big(\mu^\ell+\sum_m\sum_u \alpha_u^{\ell m}\int_0^{t_k^\ell}\beta_u e^{-\beta_u(t_k^m-s)}dN_s^m\Big) \\
&=\sum_{k=1}^{N^\ell} f^\ell(q(t_k^\ell))\mu^\ell+\sum_{k=1}^{N^\ell} f^\ell(q(t_k^\ell))\Big(\sum_m\sum_u \alpha_u^{\ell m}\int_0^{t_k^m}\beta_u e^{-\beta_u(t_k^m-s)}dN_s^m\Big)
\end{aligned} \tag{75}$$

For *term II.1*,

$$\sum_{k=1}^{N^\ell} f^\ell(q(t_k^\ell))\mu^\ell=\mu^\ell\sum_q f(q)C^\ell(q) \tag{76}$$

For *term II.2*,

$$\sum_{k=1}^{N^\ell} f^\ell(q(t_k^\ell))\Big(\sum_m\sum_u \alpha_u^{\ell m}\int_0^{t_k^m}\beta_u e^{-\beta_u(t_k^m-s)}dN_s^m\Big)=\sum_{k=1}^{N^\ell} f^\ell(q(t_k^\ell))\Big(\sum_m\sum_u \alpha_u^{\ell m} g_u^m(t_k^\ell)\Big) \tag{77}$$

The least square function defined in 20 could be calculated by summing up all these terms.

A.3.3 Gradients

Using the intermediate variables and results presented above, direct calculation shows that:

$$\begin{aligned}
\frac{\partial R}{\partial\mu^\ell}&=2\mu^\ell\sum_q D(q)f^\ell(q)^2 \\
&\quad+2\sum_q f^\ell(q)^2\Big(\sum_m \alpha_u^{\ell m}\sum_u G_u^m(q)\Big) \\
&\quad-2\sum_q f(q)C^m(q)
\end{aligned} \tag{78}$$

$$\begin{aligned}
\frac{\partial R}{\partial\alpha_u^{\ell m}}&=2\mu^\ell\sum_q f^\ell(q)^2 G_u^m(q) \\
&\quad+2\sum_{m'}\sum_{u'}\alpha_{u'}^{\ell m'}\Big(\sum_q f^\ell(q)^2 H_{uu'}^{mm'}(q)\Big) \\
&\quad-2\sum_{k=1}^{N^m} f^\ell(q(t_k^m))g_u^m(t_k^{m-})
\end{aligned} \tag{79}$$

$$\begin{aligned}\frac{\partial R}{\partial f^\ell(q)} &= \sum_q 2f^\ell(q){\mu^\ell}^2 D(q) + 4\mu^\ell f^\ell(q)\Big(\sum_m \alpha_u^{\ell m}\sum_u G_u^m(q)\Big) \\ &+ 2\sum_m\sum_{m'}\sum_u\sum_{u'}\alpha_u^{\ell m}\alpha_{u'}^{\ell m'}\Big(\sum_q f^\ell(q)H_{uu'}^{mm'}(q)\Big) \\ &- \mu^\ell C^\ell(q) - 2\sum_{k=1}^{N^\ell}\mathbb{1}_{type(t_k^{\ell\,-})=q}\sum_m\sum_u \alpha_u^{\ell m} g_u^m(t_k^\ell)\end{aligned} \tag{80}$$

References

[1] F. Abergel and Jedidi A. Long Time Behaviour of a Hawkes Process-Based Limit Order Book. *SSRN Electronic Journal*, 2015.

[2] F. Abergel, M. Anane, A. Chakraborti, A. Jedidi, and I. Muni Toke. *Limit order books*. Cambridge University Press, 2016.

[3] Frederic Abergel and Aymen Jedidi. A Mathematical Approach to Order Book Modeling. 16(5):1–40, 2010.

[4] E. Bacry, T. Jaisson, and Muzy J. Estimation of slowly decreasing Hawkes kernels: Application to high-frequency order book dynamics. *Quantitative Finance*, 16(8):1179–1201, 2016.

[5] E. Bacry, I. Mastromatteo, and J.F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 01(01):1550005, 2015.

[6] Emmanuel Bacry and Jean-François Muzy. First-and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

[7] Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys*, 2:107–144, 2005.

[8] Anirban Chakraborti, Ioane Muni Toke, Marco Patriarca, and Frédéric Abergel. Econophysics review: I. Empirical facts. *Quant. Financ.*, 11(7):991–1012, 2011.

[9] R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations Research*, 58:549–563, 2010.

[10] D. J. Daley. *An introduction to the theory of point processes: Elementary theory and methods*, volume 1. Springer, 2008.

[11] Andrew Daw and Jamol Pender. The Queue-Hawkes Process: Ephemeral Self-Excitement. *arXiv e-prints*, page arXiv:1811.04282, November 2018.

[12] Martin D Gould and Julius Bonart. Queue imbalance as a one-tick-ahead price predictor in a limit order book. *Market Microstructure and Liquidity*, 2(02):1650006, 2016.

[13] Martin D Gould, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Limit order books. *Quant. Financ.*, 13(11):1709–1742, 2013.

[14] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. "lasso and probabilistic inequalities for multivariate point-processes". *Bernoulli*, 21(1):83–143, 2015.

[15] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83—-90, 1971.

[16] W. Huang, Lehalle C.A., and Rosenbaum M. Simulating and Analyzing Order Book Data: The Queue-Reactive Model. *Journal of the American Statistical Association*, 110(509):107–122, 2015.

[17] Aymen Jedidi and Frédéric Abergel. On the stability and price scaling limit of a hawkes process-based order book model. *SSRN Electronic Journal*, pages 1–22, 05 2013.

[18] Charles-Albert Lehalle, Othmane Mounjid, and Mathieu Rosenbaum. Optimal liquidity-based trading tactics. pages 1–39, 2018.

[19] Xiaofei Lu and Frédéric Abergel. Order-book modelling and market making strategies. pages 1–23, 2018.

[20] S. Meyn and R.L. Tweedie. Stability of markovian processes iii: Foster-lyapunov criteria for continuous-time processes. *S. Advances in Applied Probability*, 25:518–548.

[21] Maxime Morariu-Patrichi and Mikko S. Pakkanen. State-dependent Hawkes processes and their application to limit order book modelling. *arXiv e-prints*, page arXiv:1809.08060, September 2018.

[22] Christine A Parlour and Duane J Seppi. Limit order markets: A survey. *Handbook of financial intermediation and banking*, 5:63–95, 2008.

[23] Marcello Rambaldi, Emmanuel Bacry, and Fabrizio Lillo. The role of volume in order book dynamics: a multivariate hawkes process analysis. *Quantitative Finance*, 17(7):999–1020, 2017.

[24] Marcello Rambaldi, Emmanuel Bacry, and Jean-François Muzy. Disentangling and quantifying market participant volatility contributions. *arXiv preprint arXiv:1807.07036*, 2018.

[25] E. Smith, J. D. Farmer, L. Gillemot, and S. Krishnamurthy. Statistical theory of the continuous double auction. *Quantitative Finance*, 6:481–514, 2003.