Modelling intensities of order flows in a limit order book

Ioane Muni Toke[*1] and Nakahiro Yoshida[†2]

[1]Laboratoire MICS, Chair Of Quantitative Finance, CentraleSupelec, Paris, France.
[1]University of New Caledonia, Noumea, New Caledonia.
[2]Graduate School of Mathematical Sciences, University of Tokyo, Tokyo, Japan.
[1,2]CREST, Japan Science and Technology Agency, Japan.

September 9, 2016

Abstract

We propose a parametric model for the simulation of limit order books. We assume that limit orders, market orders and cancellations are submitted according to point processes with state-dependent intensities. We propose new functional forms for these intensities, as well as new models for the placement of limit orders and cancellations. For cancellations, we introduce the concept of "priority index" to describe the selection of orders to be cancelled in the order book. Parameters of the model are estimated using likelihood maximization. We illustrate the performance of the model by providing extensive simulation results, with a comparison to empirical data and a standard Poisson reference.

Keywords : order book ; limit orders ; market orders ; cancellations ; state-dependent point processes ; intensity-based models.

---

[*]ioane.muni-toke@centralesupelec.fr
[†]nakahiro@ms.u-tokyo.ac.jp

1   Introduction

The limit order book is the central structure aggregating the orders of all traders to buy and sell shares of a given stock on an exchange. It is standard to simplify the complex diversity of financial messages into three types of orders : limit orders are submitted with a (limit) price into the order book, where they wait to be matched by a counterpart for a transaction ; market orders are submitted without any price and are executed immediately ; cancellations of pending limit orders is possible at any time. The order book can thus be viewed as a complex dual queueing system with price and time priority rules (see Abergel et al. (2016) for an introductory book treatment).

A partial theoretical treatment of this complex random system is possible under very simplistic assumptions, essentially assuming that the submission of limit orders, market orders and cancellations are basic Poisson processes (Cont et al. 2010, Muni Toke 2015). Exact analytical results are however limited. With appropriate scaling techniques, some limit behaviours of this complex system can be studied, see e.g. Abergel & Jedidi (2013), Huang & Rosenbaum (2015) for a price diffusion process, or Cont & De Larrard (2012) for a diffusion approximation of the volumes at the best quotes.

Another branch of study of the limit order books deals with a more statistical point of view. Smith et al. (2003) investigates the order book structure with mean field techniques. Mike & Farmer (2008) propose an empirical model of the order book that aims at reproducing some of empirical observations usually made on financial markets. Among other contributions, they propose a Student model for the placement of limit orders and a three-variable model for the cancellation of pending limit orders. The core of the submission mechanism in the order book remains however a Poisson process. Recently, Huang et al. (2015) have proposed a model in which the intensities of submission of limit, market orders and cancellations depend on the volume of the first limit. They are able to show that a queueing system with these intensities is able to reproduce some empirical features of the limit order book, such as the distribution of the first level.

In this paper we propose a general model in line with previous contributions such as Mike & Farmer (2008), Huang et al. (2015). We do not extend or specify previous models but build directly from the data. Our goal is to provide state-dependent intensities of submissions of limit and market orders that can be used for the simulation of a "realistic" limit order book. We adopt the following modelling principle : limit and market orders intensities should depend on both dimensions of the limit order book, namely the price dimension and the volume dimension. The spread is an obvious choice to include the price dimension in the modelling for both types of orders. The volume of the first level is another obvious choice for market orders, while the total volume available appears to be a good candidate for the limit orders. We define exponential forms of intensities that are convenient for two reasons: they keep the non-negativity of intensities of point processes, and they allow for a practical maximum-likelihood estimation. For the cancellation process, we introduce a new "priority index" as a main modelling variable, which turns out to be very efficient. All

proposed models are fitted on a database of 10 consecutive trading days (January 17th-28th, 2011) for six different liquid stocks traded on the Paris stock exchange during the construction phase of the model, then on a much longer sample (2011-2013) to test its robustness.

The rest of the paper is organized as follows. Section 2 briefly describes the data and its preparation. Section 3 provides empirical insights on the intensity of submission of market orders and build a convenient parametric model. Section 4 introduces a similar model for the intensities of limit orders and provide a very flexible Gaussian mixture model for the placement of limit orders, that is able to reproduce the multi-modality of the empirical distribution. Section 5 shows that the "priority volume", i.e. the volume standing in front of a pending order according to time-price priority rules is a good candidate for the modelling of the "placement" of cancellations. Section 6 provides insights on the stability of the model across time by fitting it weekly on a two-year-and-a-half-long sample. Finally, Section 7 develops a market simulator and provides extensive results of simulations of our model fitted on market data. The performances of the simulations are analysed, in particular with respect to a standard Poisson reference.

2 Data

We use data extracted from the Thomson-Reuters Tick History (TRTH) database. We randomly select six liquid stocks from the CAC 40 index (i.e. stocks among the highest capitalizations exchanged at the Paris Bourse) : Air Liquide (Reuters Instrument Code (RIC): AIRP.PA), Alstom (ALSO.PA), BNP Paribas (BNPP.PA), Bouygues (BOUY.PA), Carrefour (CARR.PA), Electricite de France (EDF.PA). These stocks represent a wide panel of liquidity for CAC 40 index: BNPP.PA is a heavily traded stock, one of the most traded on the Paris Stock Exchange, while EDF.PA is less actively traded and is a much smaller capitalization (EDF.PA has even since been removed from the CAC 40 index on December 21st, 2015).

For each stock, two files can be extracted from the TRTH database, which are commonly called the trades file and the quotes file. The quotes file is a sequence of snapshots of the limit order book, listing all the modifications due to the processing of orders, each modification being timestamped with a millisecond resolution. This file can be parsed to extract a preliminary order flow of limit orders (increase of the available liquidity on a given side at a given price) and cancel orders (decrease of available liquidity on a given side at a given price). The trades file is then parsed and matched to the previous (preliminary) order flow to identify and convert some of the cancel orders into market orders.

For each trading day, we keep the subset of limit orders, market orders and cancellations occurring between 9:05 in the morning and 17:25 in the afternoon, i.e. we keep the whole trading day except the first five minutes of the day, following the opening auction, and the last fives minutes of the day preceding the closing auction. The data in these very active periods seems indeed of a lesser quality and not always reliable. In order to build the model in Sections 3, 4 and 5, we

keep only orders occurring on the ask side of the limit order book, and we glue the ten days of order flows as in an artificial continuous sample. The same procedure is applied in Section 6 for some stability testing (weekly glueing of the ask side, on a longer sample). Finally, in order to run simulations of the model in Section 7 however, we will adopt a more practical point of view and use both sides of the book but only one day of trading at a time to fit and test the model, without any glueing of consecutive trading days.

As a result of this process of data preparation, we have for each stock and each period (one trading day or glued trading days) a list of orders (the order flow) and for each order a list of variables describing the limit order book at the time of submission : spread, volume at the best quotes, total liquidity available at the ten best quotes. Details and performances of the data preparation procedure can be found in Muni Toke (2016). As a final step specific to this paper, since we are going to deal with simple point processes — for which the probability of occurrences of simultaneous events is zero — we combine orders of the same type and the same timestamp (precise to the millisecond) into a single order with a size equal to the sum of the sizes of the combined orders.

Let us add a few words on the units of this data. Time is measured in seconds (with a millisecond precision). Prices in the order book must be integer multiples of a tick size which is fixed by the exchange. In our sample, AIRP.PA and BNPP.PA have a 0.01 EUR ticksize, while the four other stocks have a 0.005 EUR ticksize. As for the volumes, they are numbers of shares. For ease of computations and presentation, all volumes are normalized by a stock-dependent quantity equal to the median of the trade size (MTS) for this stock. In order to keep this volumes integers, we round the results to the smallest larger integer (ceiling). As a result, 0 means really no share, while 1 is a small non-zero volume. These remarks should explain the x-axis scales of the graphs of the following sections.

Remark 1. By normalizing the volumes, we are in line with e.g. Cont et al. (2010) (normalization using the average size of limit orders) or Huang et al. (2015) (normalization using the average event size per limit). This normalization however leads inevitably to a loss of information. By not normalizing the sizes, we would observe peaks in the distribution for "round" values (50, 100, 200, etc.), as reported long ago by Challet & Stinchcombe (2001). The median trade size (MTS) we use here for normalization is a bit finer than the average trade size (ATS) as it is less sensitive to a few extremely high volumes that are exceptionally observed but are not the focus of this model. To get a better grasp of the empirical distribution, one may be tempted to use smaller bin sizes, but this will quickly make the distribution noisier, and finding the "visually optimal" binning is an open problem.

Finally, after all these steps of data preparation, the resulting order flow can be summarized by a few average daily statistics given in Table 1. BNPP.PA is the most traded stock of the sample, both in number of orders and volumes. It has the smallest average spread (in ticks) and the highest

4

RIC	# Limit	# Market	# Cancel	MTS	Traded Volume	q_1	spread
AIRP.PA	136,953	4,799	128,761	64.8	527,707	4.13	2.80
ALSO.PA	116,564	6,432	106,003	135.2	1,728,357	4.95	3.65
BNPP.PA	188,272	10,140	169,262	175.8	3,609,045	5.98	2.01
BOUY.PA	78,727	3,844	72,652	154.5	1,065,902	4.57	2.93
CARR.PA	96,800	4,337	88,689	208.0	1,639,336	4.73	2.44
EDF.PA	80,422	4,168	74,202	145.1	1,098,381	5.26	2.27

Table 1: Average daily statistics describing the order flow for the six stocks considered, between January, 17th and January, 28th, 2011. q_1 is the size of the first limit of the order book, expressed in MTS. Spread are expressed in number of ticks.
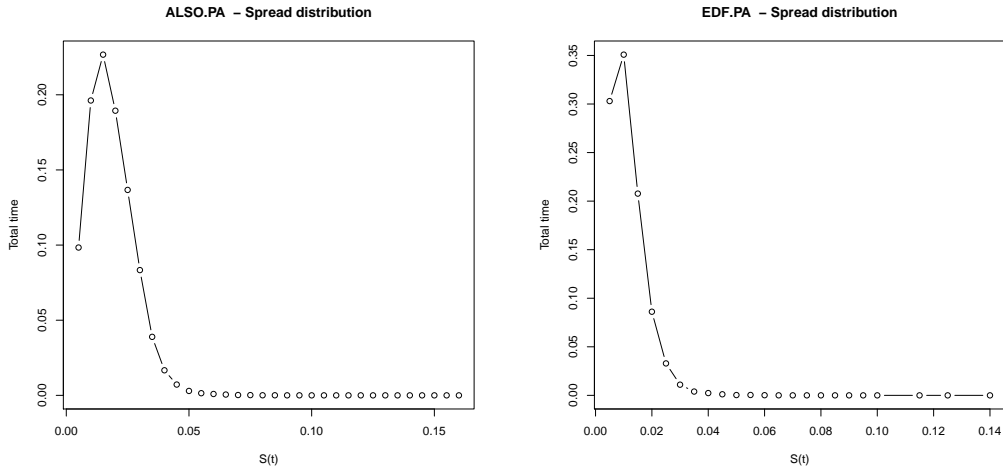

Figure 1: Empirical distribution of the spread for two stocks, ALSO.PA (left) and EDF.PA(right).

average volume available at the best quote q_1 (in MTS). ALSO.PA is the second most actively traded stock and is singular by its large average spread. AIRP.PA, BOUY.PA, and CARR.PA are quite comparable in activity and average values. EDF.PA is singular by its larger average volume available at the best quote.

Quantities in Table 1 are daily averages of normalized values. To give the reader a better view of the data, we plot on Figures 1 and 2 the empirical distribution of the spread (in EUR) and the empirical distribution of the volume at the best quote q_1 (in shares), for two stocks of our sample.

These graphs show the dominance of a few low values, and a long tail of distribution. These observations will help assessing the performances of the models proposed in the following sections.
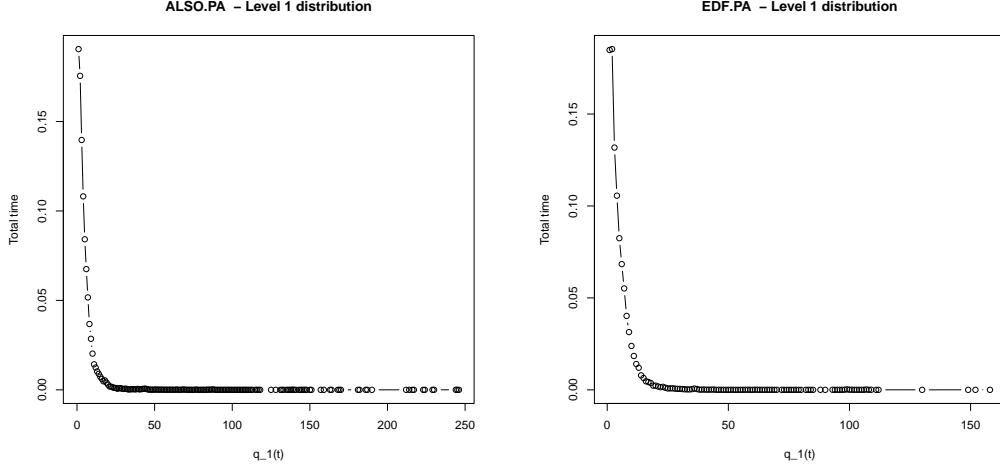
Figure 2: Empirical distribution of q_1 for two stocks, ALSO.PA (left) and EDF.PA(right).

3   Market orders

Let N^M be the point process of submission of market orders in the limit order book and let λ^M be its instantaneous intensity. Our goal is to identify a simple parametric model for λ^M, which should be based on meaningful variables and be easy to estimate. We therefore identify two covariates to model λ^M : the spread S and the volume at the best quote q_1 (on the side of submission).

Let us first investigate the spread. Using common financial knowledge, one should expect specific variations of the intensity as a function. Firstly, λ^M should be decreasing with S. Indeed, if a trader needs to buy a share when S is equal to one tick, he cannot gain priority in the limit order book, and therefore has to submit a buy market order to be the first to buy the best quote. On the contrary, if the spread is large, it is sufficient to submit a buy limit order just above the best bid quote to be the first in line for the next sell-initiated transaction.

We compute on our samples an estimator of the spread-dependent intensity of limit orders:

$$\hat{\lambda}^M(S) = \frac{N^M(S)}{T(S)}, \tag{1}$$

where $N^M(S)$ is the total number of market orders submitted when the spread is equal to S and $T(S)$ is the total time during which the spread is equal to S in the sample. As an illustration, $\hat{\lambda}^M(S)$ is plotted in Figure 3 (left panel) for one of the stocks of the sample (more results will be given below). As expected, the intensity of submission of market orders is decreasing with the spread. However, this decrease does not to go to zero, and even seem to increase slightly for very large values of the spread. A plausible interpretation is that when the spread increases above usual levels, this may indicate a highly volatile period with many orders submitted. Subsequent
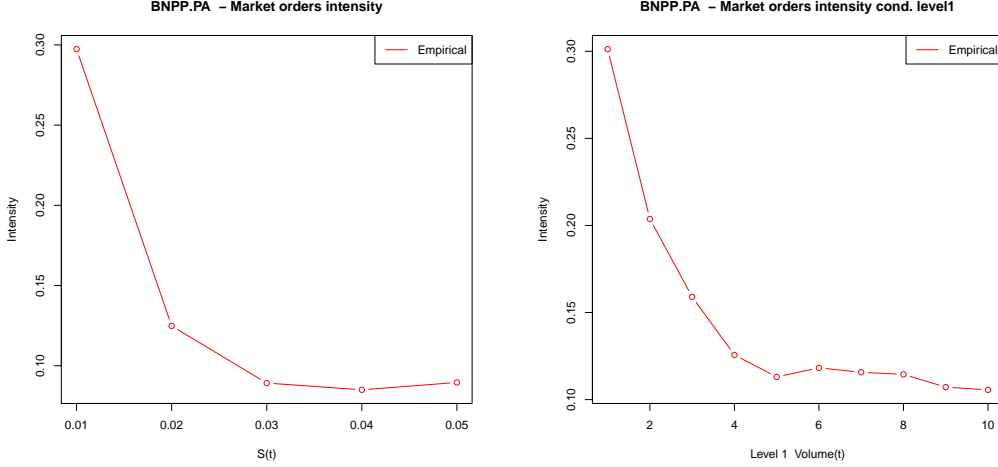
Figure 3: Left panel : Empirical λ^M as a function of the spread $(\hat{\lambda}^M(S))$. Right panel : Empirical λ^M as a function of the spread $(\hat{\lambda}^M(q_1))$. Since data may be noisy for very high values of the parameters, the x-axes span 99% of the empirical distribution of S and 90% of the empirical distribution of q_1.

uncertainty might translate into a "rush" for liquidity maintaining $\hat{\lambda}^M(S)$ above zero.

Following these empirical results, one might propose the following parametric model to express the functional dependence of λ^M on S :

$$\lambda^M(S) = \exp\left(\beta_0 + \beta_1 \ln(S) + \beta_{11}[\ln(S)]^2\right). \tag{2}$$

The exponential form ensure that λ^M remains non-negative. The quadratic argument allows the non-monotony of λ^M instead of the power-law form obtain with only one term. The preference for the logarithm of the spread instead of the spread itself is detailed in Remark 4.

Let us now turn to the second explaining variable considered here, the volume q_1 of the best quote (on the side of submission, ask for a buy market order, bid for a sell market order). We compute on our samples an estimator of the q_1-dependent intensity of limit orders:

$$\hat{\lambda}^M(q_1) = \frac{N^M(q_1)}{T(q_1)}, \tag{3}$$

where $N^M(q_1)$ is the total number of market orders submitted when the volume on the (same side) best quote is equal to q_1 and $T(q_1)$ is the total time during which this volume is equal to q_1 in the sample. Recall that the unit for q_1 is the median of the trades sizes. Results are plotted on Figure 3 (right panel). One observes that λ^M increases as q_1 decreases, as expected. Indeed, when q_1 is small, the probability that the first limit vanishes increases. This is an incentive for traders to

grab the last shares available at the current price, leading to a "rush for liquidity". This monotony however is justified for small values of q_1 and there is no obvious reason that the intensity should go to zero for large values of q_1. Figure 3 suggests that we can use a functional dependency on q_1 similar to the one suggested for the spread :

$$\lambda^M(q_1) = \exp\left(\beta_0 + \beta_2 \ln(1+q_1) + \beta_{22}[\ln(1+q_1)]^2\right). \tag{4}$$

We can finally combine the two dependencies into one single model and add an potential interaction term between the two covariates. We thus obtain the following parametric model for the intensity of submission of market orders in a limit order book:

$$\lambda^M(t; S(t), q_1(t)) = \exp\left[\beta_0 + \beta_1 \ln(S(t)) + \beta_{11}[\ln(S(t))]^2 + \beta_2 \ln(1+q_1(t)) + \beta_{22}[\ln(1+q_1(t))]^2\right.$$
$$\left. + \beta_{12} \ln(S(t)) \ln(1+q_1(t))\right]. \tag{5}$$

This model can be estimated by likelihood maximization. To emphasize the dependency on the the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_{11}, \beta_2, \beta_{22}, \beta_{12})$ to be fitted, we write $\lambda^M(t; S(t), q_1(t)) = \lambda^M(t; \boldsymbol{\beta})$ when dealing with the estimation. Log-likelihood $\mathcal{L}_T^M$ for the point process $\{N^M(t), t \in [0, T]\}$ as a function of the parameter vector $\boldsymbol{\beta}$ is defined as:

$$\mathcal{L}_T^M(\boldsymbol{\beta}) = \int_0^T \ln\left(\lambda^M(t; \boldsymbol{\beta})\right)\, dN_t^M - \int_0^T \lambda^M(t; \boldsymbol{\beta})\, dt. \tag{6}$$

Let $\{t_i^M\}$ be the set of arrival times of market orders in our sample, $\{t_i^S\}$ the set of times of jumps of the spread process S, $\{t_i^{q_1}\}$ the set of times of jumps of the first limit process q_1. Then the log-likelihood on the sample is numerically computed as follows:

$$\mathcal{L}_T(\boldsymbol{\beta}) = \beta_0 N^M(T) + \beta_1 \sum_{t_i^M} \ln S(t_i^M -) + \beta_{11} \sum_{t_i^M}[\ln S(t_i^M -)]^2$$
$$+ \beta_2 \sum_{t_i^M} \ln(1 + q_1(t_i^M -)) + \beta_{22} \sum_{t_i^M}[\ln(1 + q_1(t_i^M -))]^2 + \beta_{12} \sum_{t_i^M} \ln S(t_i^M -) \ln q_1(t_i^M -)$$
$$- \sum_{t_i \in \{t_i^S\} \cup \{t_i^{q_1}\}} \exp\left[\beta_0 + \beta_1 \ln S(t_i -) + \beta_{11}[\ln S(t_i -)]^2\right.$$
$$\left. + \beta_2 \ln(1 + q_1(t_i -)) + \beta_{22}[\ln(1 + q_1(t_i -))]^2 + \beta_{12} \ln S(t_i -) \ln(1 + q_1(t_i -))\right](t_i - t_{i-1}). \tag{7}$$

It is then numerically maximized using the routine `mle2` of the `bbmle` package in the `R` language. Results for all the stocks of our samples are given in Table 2. For simplicity of presentation these results are shown for the ask side only (buy market orders), but results for the bid side are similar.

RIC	β_0	(std)	β_1	(std)	β_{11}	(std)	β_2	(std)	β_{22}	(std)	β_{12}	(std)
AIRP.PA	-0.527	0.389	1.730	0.190	0.370	0.023	-1.080	0.091	0.203	0.014	-0.016	0.019
ALSO.PA	6.193	0.299	4.034	0.131	0.537	0.014	-1.770	0.061	0.235	0.007	-0.125	0.013
BNPP.PA	3.713	0.452	3.100	0.220	0.482	0.027	-1.463	0.058	0.160	0.003	-0.126	0.013
BOUY.PA	1.426	0.532	2.698	0.224	0.425	0.024	-0.734	0.098	0.155	0.010	0.015	0.021
CARR.PA	-0.694	0.678	1.565	0.281	0.298	0.029	-0.723	0.114	0.251	0.012	0.111	0.021
EDF.PA	7.863	0.646	5.066	0.269	0.629	0.028	-1.486	0.118	0.154	0.013	-0.134	0.022

Table 2: Maximum-likelihood parameters for the intensity λ^M for all the stocks of our sample.

Table 2 provides the numerical values of the parameters as well as the standard deviation estimated by the maximization routine. These standard deviations assess the quality of the fitting and verify that all the fitted values are significant to a high-level, except for small β_0's for AIRP.PA and CARR.PA, and small β_{12} (AIRP.PA and BOUY.PA). This last fact concerning β_{12} is not very surprising as the joint distribution between S and q_1 is quite difficult to characterize, and an independence hypothesis between these two modelling variables might not be unreasonable for some stocks.

We now provide several graphs to illustrate the fitting performance of the model. We first plot the empirical intensity as a function of the spread ($\hat{\lambda}^M(S)$) and the "marginal" spread-dependent intensity $\tilde{\lambda}^M(S)$ computed by our model. This "marginal" represents the dependence on the spread when q_1 is distributed as in the sample, i.e. it is computed with obvious notations as :

$$\tilde{\lambda}^M(S) = \sum_q \lambda^M(t; S, q)\mathbf{P}(q_1 = q). \tag{8}$$

Similarly, we then plot for each stock the empirical intensity as a function of the level q_1 ($\hat{\lambda}^M(q_1)$) and the "marginal" q_1-dependent intensity $\tilde{\lambda}^M(q_1)$ computed by our model as :

$$\tilde{\lambda}^M(q_1) = \sum_s \lambda^M(t; s, q_1)\mathbf{P}(S = s). \tag{9}$$

Results are given on Figure 4 for the dependence on the spread and on Figure 5 for the dependence on the volume at the best quote q_1. For the sake of conciseness, we reproduce only two "representative" stocks (see the discussion below). The "marginal" intensities allow for a synthetic view of the modelling intensity. In order to provide the reader with the full view of the fitting, we finally plot the spread-dependent empirical intensities given the volume q_1, and symmetrically the q_1-dependent intensities given the spread S. Results are plotted on Figures 6 and 7, for two representative stock and each time for the first five most probables occurrences of the variables. There again, two "representative" stocks are chosen.

Let us start with Figure 4. It turns out that the marginal fitting for the spread is always good, and even excellent for most stocks. Results for AIRP.PA, BNPP.PA, CARR.PA (not shown) are
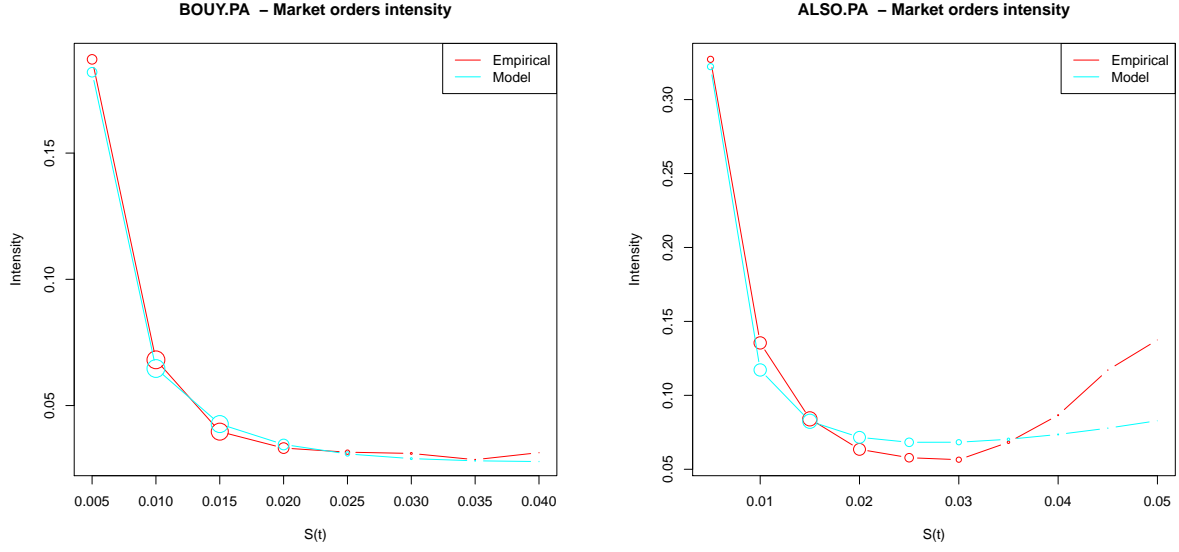
Figure 4: Empirical $(\hat{\lambda}^M(S))$ and model $(\tilde{\lambda}^M(S))$ intensities of market orders as functions of the spread S. Dots sizes are proportional to the empirical frequency of the spread. x-axis spans 99% of the spread distribution.
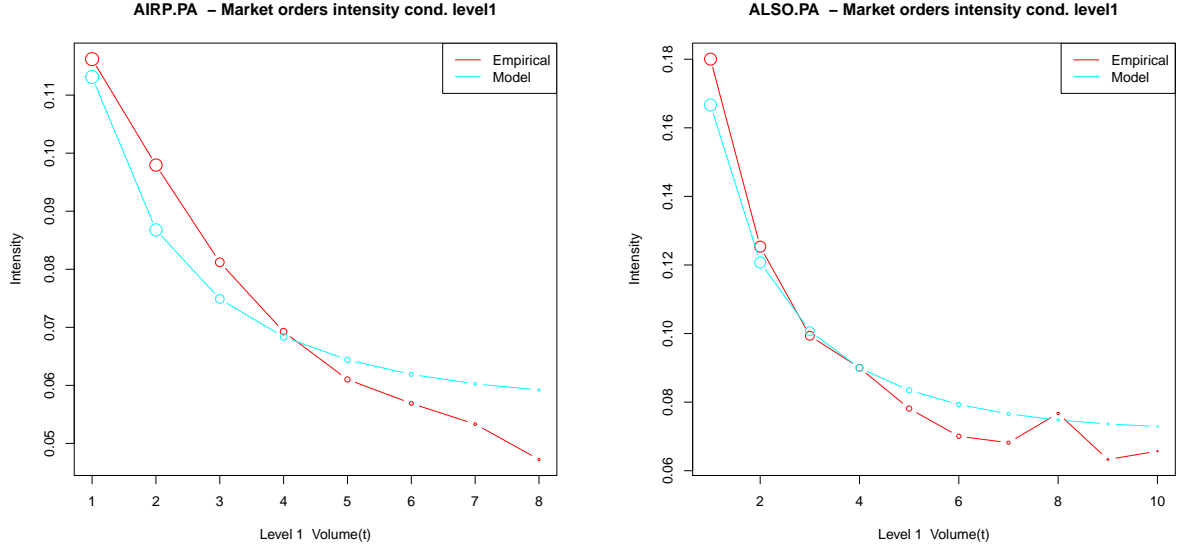

Figure 5: Empirical $(\hat{\lambda}^M(q_1))$ and model $(\tilde{\lambda}^M(q_1))$ intensities of market orders as functions of the volume of the first limit q_1. Dots sizes are proportional to the empirical frequency of q_1. x-axis spans 90% of the distribution of q_1.
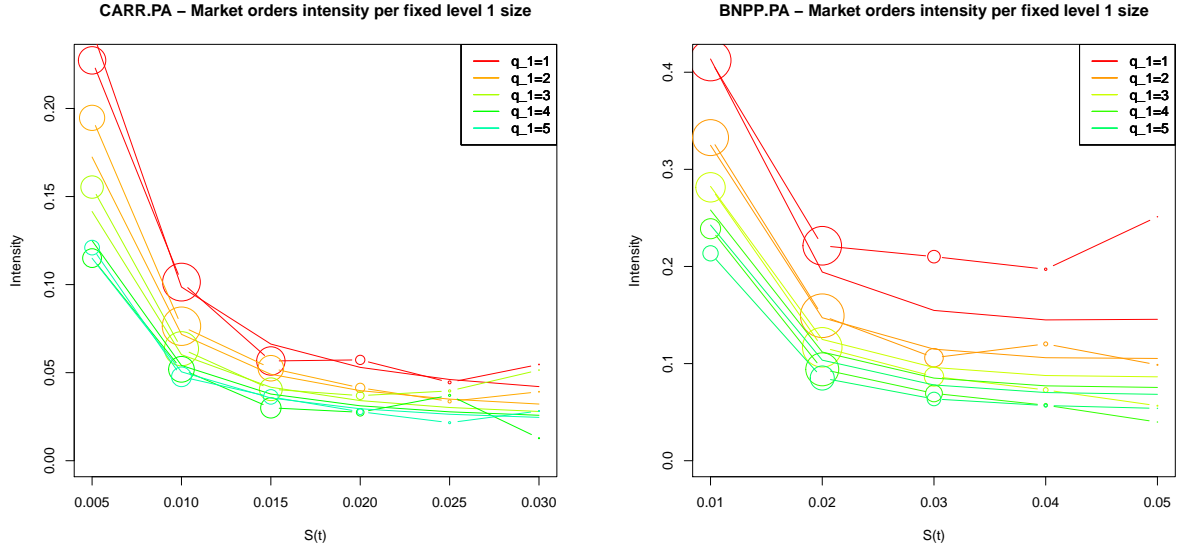
10

Figure 6: q_1-conditional intensities as functions of the spread. Lines with large dots represent empirical intensities and solid lines the fitted model intensities. Each q_1 level has one color. Dots sizes are proportional to the empirical frequency of the state (S, q_1).
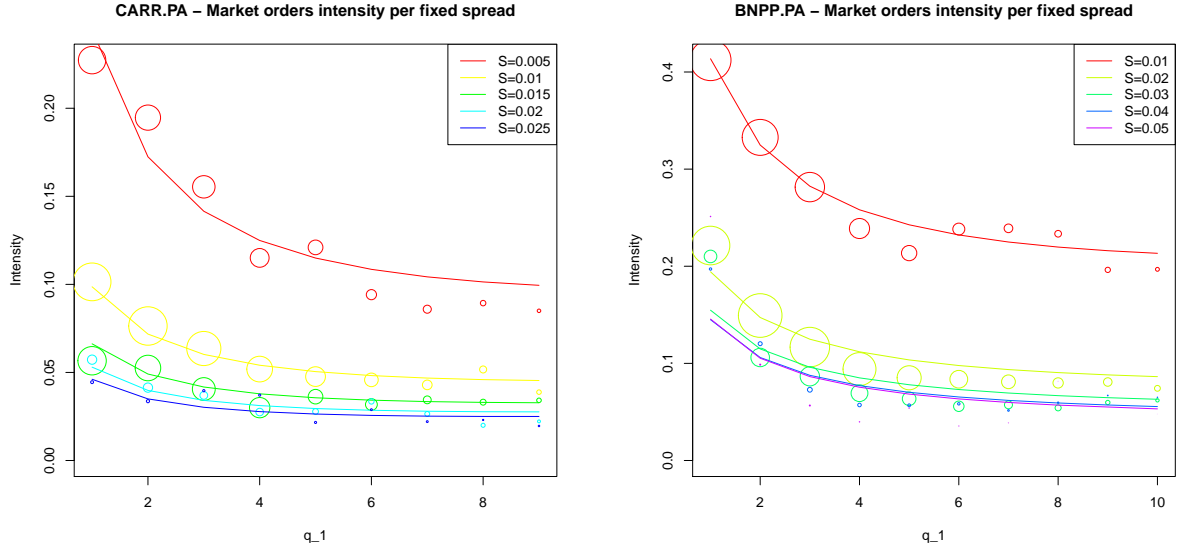

Figure 7: Spread-conditional intensities as functions of q_1. Dots represent empirical intensities and lines the fitted model intensities. Each spread level has one color. Dots sizes are proportional to the empirical frequency of the state (S, q_1).

11

all similar to BOUY.PA (left panel), showing an excellent visual fit. It seems that the model fails to catch the full extent of the increase of λ^M observed for large values of the spread for some stocks (ALSO.PA (right panel), and to a lesser extent EDF.PA (not shown)). It is however important to recall that high-spread values are very rare events, as was observed on Figure 1. This shows for example that the points that are less well fitted — e.g. the rightmost points of the graphs — actually represent a very small part of the spread distribution. It is therefore perfectly normal that the MLE estimation favours the main part of the distribution (left part of the graphs). The sizes of the dots on the figure are proportional to the observed spread frequency to emphasize this point. This good fitting with respect to the spread is confirmed on Figure 7 where each spread-conditional intensity is well modelled for each stock. The results for two stocks are shown but all results are comparable, showing a good fit for frequent events.

Continuing the analysis of the graphs, we observe on Figure 5 that the quality of the fitting of the dependency on the volume q_1 seems a bit poorer. Two stocks are shown as examples. On the one hand, AIRP.PA (left panel) shows a quite regular decrease, with a curvature smaller than the one proposed by the model. BOUY.PA, CARR.PA and EDF.PA (not shown) are similar. On the other hand, ALSO.PA (right panel) and BNPP.PA (not shown) have higher intensities for lower values of q_1 and decrease faster, which the model can eventually capture. Figure 6 plots the conditional intensity given q_1. Again we show only two representative graphs for brevity. CARR.PA (left panel) is a good fit, in which all states are well fitted (AIRP.PA and BOUY.PA are similar). EDF.PA and ALSO.PA (not shown) are also very well fitted with a lesser quality for extreme spread values, as expected from the discussion of Figure 4. BNPP.PA (right panel) is a good fit except for simultaneous large spread and small volume at the best quote ($S \geq 3$ ticks and $q_1 = 1$), which intuitively is a situation of an unusually high level of activity. Figure 6 clearly shows that larger levels of q_1 have less influence, leading to the collapsing of the conditional intensities on the same curve. The secondary role of larger values of q_1 is not surprising. There again, recall the examples of empirical distributions of q_1 given in Figure 2. The body of the distribution is clearly to the left, leaving less weight for the higher values.

Therefore, the proposed model is overall a good fit, especially if we keep in mind that despite their differences we have managed to propose the same functional form for the dependence on the spread and the dependence on the volume at the best quote q_1. We end this section by three modelling remarks, opening potential future works, and then move on to modelling of limit orders.

Remark 2. The form $\ln(1 + q_1(t))$ is here preferred to $\ln(q_1(t))$ for flexibility as it allows for a normalized volume q_1 equal to zero. This is not the case in this paper since we have rounded above normalized volumes, so that 0 is really 0, not a small volume. But the difference being marginal, we keep the general (right-shifted) form.

Remark 3. The likelihood analysis here is a conditional likelihood analysis given $S(t)$ and $q_1(t)$, or a regression analysis with these explanatory variables. We discuss the modeling of limit orders

and cancellations in the following sections, where a certain parametric model is introduced for each order. Naturally, these models should be unified to describe the whole picture of all orders though we do not pursue the integration of models in this paper.

Remark 4. In the above construction of a model for the intensity, the exponential of a quadratic form of the logarithm of the variable is selected by the AIC criterion over an exponential of a quadratic form of the natural variable. Hence our choice that may not appear standard at first sight. Furthermore, significance of every parameter suggests that we could introduce more explanatory variables and select a suitable model by a certain information criterion or a sparse estimation method. This is future work.

4   Limit orders

We now turn to the modelling of limit orders. Defining a limit order requires one dimension more than defining a market order : its (limit) price has to be chosen upon submission. We have decided to treat the two problems separately. In a first subsection 4.1, we deal with the point process N^L counting all limit orders (at any prices), with an instantaneous intensity λ^L. The distribution of prices is assumed to be independently defined and will be discussed in the following subsection 4.2.

4.1   Modelling limit orders intensities

Similarly to what we did for market orders, we choose two variables for our modelling. The price dimension is represented by the spread S. As for the "volume" dimension, we investigate the total volume available in the limit order book at the side of submission (more precisely the sum of all the liquidity available up to the tenth limit), denoted here Q_{10}. Since λ^L deals with all limit orders, Q_{10} appears obviously more relevant that q_1 as a modelling variable.

Following our modelling principles, we propose the following model for limit orders :

$$\lambda^L(t; S(t), Q_{10}(t)) = \exp\left[\beta_0 + \beta_1 \ln(S(t)) + \beta_{11}[\ln(S(t))]^2 + \beta_2 \ln(1 + Q_{10}) + \beta_{22}[\ln(1 + Q_{10})]^2\right.$$
$$\left. + \beta_{12} \ln(S(t)) \ln(1 + Q_{10})\right]. \tag{10}$$

Here, we expect the intensity λ^L to increase with the spread (by an argument exactly symmetric to the one we have used in Section 3, see above). We also expect it to increase when Q_{10} decreases since by an expected stability mechanism, a global drop in the available volume should be an incentive to provide more liquidity. As mentioned before, these monotonous variations guessed by "common financial sense" are only expected to be observed for frequent values of the modelling variables, since (rare) extreme values of the parameter are noisy and therefore difficult to characterize.

The model defined at Equation (10) can be fitted by maximization of the likelihood. It is

RIC	β_0	(std)	β_1	(std)	β_{11}	(std)	β_2	(std)	β_{22}	(std)	β_{12}	(std)
AIRP.PA	5.772	*0.108*	-0.028	*0.031*	0.048	*0.004*	-2.182	*0.041*	0.263	*0.005*	0.079	*0.005*
ALSO.PA	14.516	*0.102*	3.964	*0.030*	0.443	*0.003*	-2.386	*0.035*	0.248	*0.004*	-0.024	*0.005*
BNPP.PA	6.472	*0.087*	1.898	*0.040*	0.285	*0.005*	-0.645	*0.018*	0.101	*0.001*	0.084	*0.004*
BOUY.PA	17.042	*0.128*	3.090	*0.041*	0.296	*0.004*	-4.789	*0.040*	0.507	*0.005*	-0.113	*0.007*
CARR.PA	12.223	*0.150*	0.699	*0.046*	0.116	*0.005*	-4.635	*0.048*	0.563	*0.005*	0.079	*0.007*
EDF.PA	15.176	*0.155*	1.971	*0.047*	0.164	*0.005*	-4.443	*0.052*	0.456	*0.006*	-0.064	*0.007*

Table 3: Fitted coefficients by maximum likelihood estimation for the intensity of limit orders.

straightforward to modify the formula given at Equation (7) to obtain the log-likelihood of the model, so we skip it for brevity. The numerical results of the maximum likelihood estimation are given in Table 4.1. There again, standard deviations are provided to assess the quality of the fitting.

We now provide graphical illustration of the quality of the fitting of the model. One can straightforwardly adapt Equations (8) and (9) to compute the "marginal" intensities of limit orders with respect to the spread and Q_{10}. These are plotted on Figures 8 and 9 where they are compared to the empirical intensities. As for the dependence on the spread, we observe that the intensity $\hat{\lambda}^L(S)$ exhibits several shapes. There is indeed an increase for large spreads, as we expected, but for small spread values we observe either a strong decrease (BOUY.PA, top right panel), a slower decrease (BNPP.PA, top left panel) or an immediate (slow) increase (EDF.PA, bottom right panel). The interesting point is that the proposed model is flexible enough to reproduce all observed shapes, except when the empirical distribution shows a very irregular behaviour (unexpected drop for large spreads for AIRP.PA, bottom left panel). Note that one could probably get better fits (for the eye) with some least-squares regression techniques, but the maximum-likelihood estimation chosen here emphasizes on the main body of the distribution, i.e. small spreads.

As for the dependence on the total volume available in the book on the side of submission Q_{10}, we observe that the intensity increases when the available liquidity decreases, as expected. The model is able to reproduce this feature : AIRP.PA, ALSO.PA and EDF.PA (not shown) are similar to BOUY.PA (right panel). Again the fitting procedure favours the main part of the distribution (BNPP.PA, left panel), but the model is however able to grasp an increase of the intensity when Q_{10} increases above average (observed on CARR.PA, not shown). Note that the empirical part of the plot is quite noisy : this is because we have kept for coherence the binning size equal to the median trade size (MTS), which is small compared to the order of magnitude of the whole volume available in the limit order book.

Finally, the proposed model is once again a good fit, especially if we keep in mind that we have managed to propose the same functional form for the limit and market orders intensity, including both a price and a volume variable, following our modelling principle.
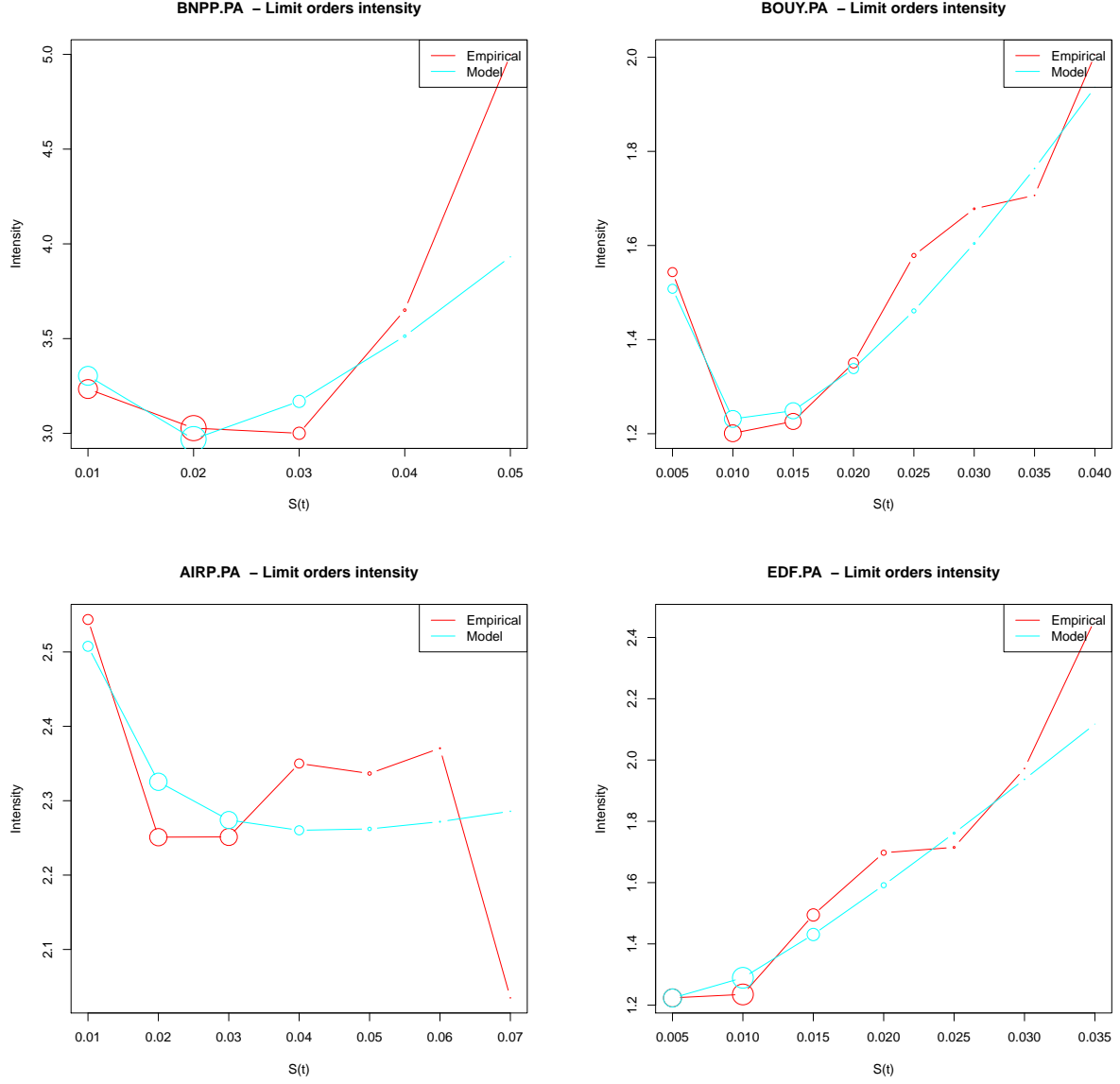
Figure 8: Empirical $(\hat{\lambda}^L(S))$ and model $(\tilde{\lambda}^L(S))$ intensities for limit orders as functions of the spread S.
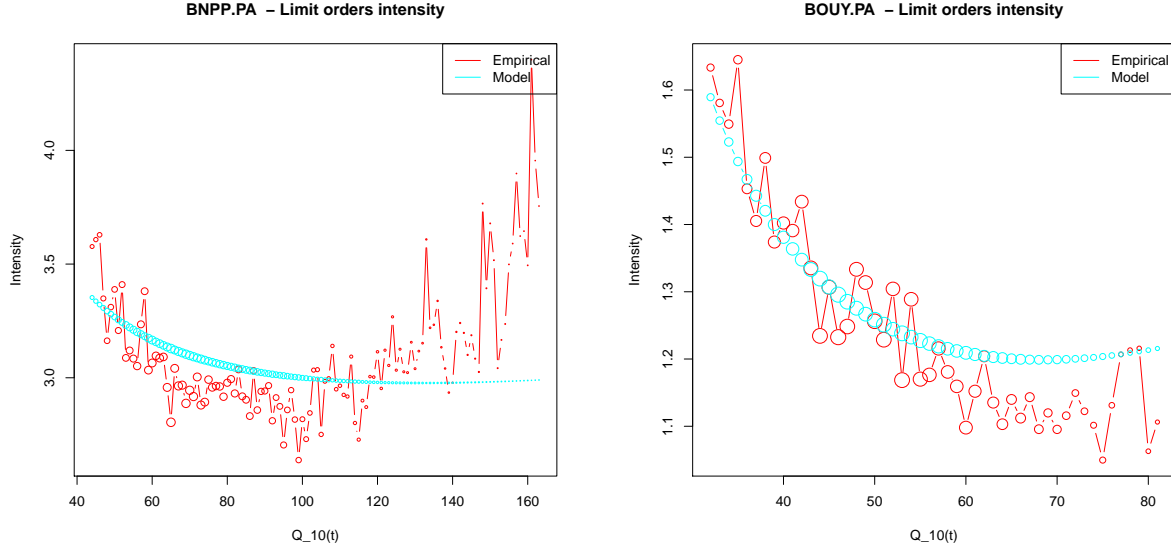
Figure 9: Empirical ($\hat{\lambda}^L(Q_{10})$) and model ($\tilde{\lambda}^L(Q_{10})$) intensities for limit orders as functions of the total book volume Q_{10}.

4.2 Modelling the placement of limit orders

Modelling the placement of limit orders can a be difficult challenge. The support of any placement distribution is indeed state-dependent : in our model that distinguishes between three types of orders (limit, market, cancellation), one cannot submit a sell/buy limit order below/above the current best bid/ask. Such an order should be a market order.

With a simulation perspective, one can settle for a general distribution and then drop at the time of simulation any non-acceptable price (see Section 7). Using this technique, Mike & Farmer (2008) argued that the Student distribution centred around the current best quote is a good fit for the placement of limit orders (using data for the stock AstraZeneca on the London Stock Exchange). In the same spirit, we will use continuous distributions on $\mathbb{R}$ to model the placement. 0 will be the current best quote. We consider the placement distribution as a function on the *continuous* variable price, and then integrate this density to obtain the discrete probability distribution of the placement of limit orders on the grid of integers numbers of ticksize. If $\pi^L : \mathbb{R} \to \mathbb{R}_+$ is the continuous density of placement of limit orders and δ is the ticksize, then $\int_{(n-0.5)\delta}^{(n+0.5)\delta} \pi^L(u)\, du$ is the probability that the limit order is submitted at price $p = n\delta$.

We propose here two models. The first one is a generalized version of the Mike & Farmer (2008) proposition in which the limit orders are placed according to a location-scale version of the Student

16

RIC	π_1	π_2	π_3	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
AIRP.PA	0.211	0.309	0.480	0.050	2.751	4.849	0.793	1.249	2.903
ALSO.PA	0.191	0.288	0.522	0.064	3.056	4.702	0.726	1.659	3.371
BNPP.PA	0.319	0.304	0.378	0.192	1.905	4.570	0.719	0.935	2.733
BOUY.PA	0.193	0.295	0.512	0.004	2.886	4.583	0.704	1.418	3.145
CARR.PA	0.235	0.283	0.483	0.159	2.405	4.560	0.768	1.214	2.695
EDF.PA	0.242	0.269	0.489	0.024	2.558	4.400	0.695	1.158	2.865

Table 4: Fitted parameters for the normal mixture model for the placement of limit orders. Parameters of the Gaussian distributions are expressed in number of ticksizes.

distribution:

$$\pi^L(p; \mu, \sigma, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{1}{\nu} \left(\frac{x-\mu}{\sigma} \right)^2 \right)^{-\frac{\nu+1}{2}}. \tag{11}$$

This model is interesting as it has only three parameters. However empirical data suggests that for some of the stocks we have studied placement of limit orders is often multi-modal. To our knowledge this observation has not been made before. One indeed observes a peak of submission at the best quote, and then another mode inside the book, a few ticks away from the best quote. In order to reproduce this complex distribution we use a mixture of $G = 3$ normal distributions:

$$\pi^L(p; G, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{G} \pi_i \phi(p; \mu_i, \sigma_i), \tag{12}$$

where $\phi(\bullet; \mu, \sigma)$ is the density of the Gaussian distribution with parameters (μ, σ).

The normal mixture model is fitted with the mclust package of the R language. The fitted parameters are given in Table 4. For all stocks, the fitted mixture model exhibits the same components. One Gaussian is centred on the best quote and very thin (standard deviation of two-third of a ticksize). This distribution accounts for roughly 20-25% of the submitted limit orders, and helps modelling the peak of limit orders submitted at the best quote. Two other Gaussian distributions are further away in the book (roughly 2-3 and 4-5 ticks away from the best quote), and help modelling the second mode observed and the more passive limit orders.

In order to illustrate the quality of the fitting obtained, Figure 10 plots the model distribution compared to the empirical one for two stocks of the sample. AIRP.PA, ALSO.PA, CARR.PA and EDF.PA (not shown) are all similar to BOUY.PA (left panel). They all show a bimodal empirical distribution that is very well grasped by the mixture model. BNPP.PA (right panel) is singular in that it is mono-modal but with a large hump on the right side, that is also well fitted by the mixture model. The fitted location-scale Student is given for comparison. This mono-modal distribution is in our sample centred on the maximum inside the book, a few ticks away from the best. As a result, it underestimates on the one hand the number of orders submitted at the best quote, but
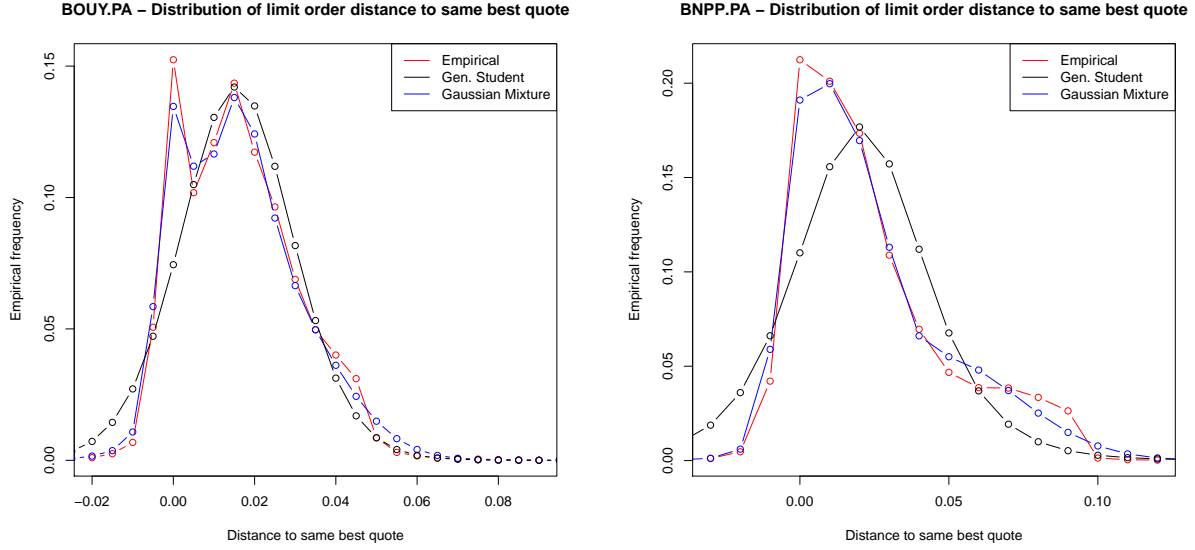
Figure 10: Empirical and model distribution of the placement of limit orders.

on the other hand it overestimates the number of aggressive orders submitted inside the spread.

Remark 5. We observe that the multi-modality of the placement of limit orders strongly depends on the observed spread. It is usually stronger for small spreads, and disappears for larger spread. This can be interpreted as follows. When the spread is smaller than usual, the market participants anticipate its widening, thus providing liquidity a few ticks inside the book besides the usual liquidity provided at the best quote. Hence the appearance of two peaks in the distribution on the placement of limit orders, and the strong multi-modality. When the spread is large, market participants anticipate its tightening, thus providing more liquidity close to the best quote, hence the disappearance of the multi-modality.

It is easy to generalize our model given at Equation (12) to a spread-dependent model, by splitting our sample according to the observed spread and then fitting spread-dependent parameters:

$$\pi^L(p, S; G, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{G} \pi_i(S)\phi(p; \mu_i(S), \sigma_i(S)). \tag{13}$$

This would increase the number of parameters of the model but allow for a better flexibility in the modelling of the placement of limit orders. With the simulation of Section 7 in mind and given the good performances of the proposed fit, we stick, at least for now, to the unconditional model.

5   Cancellations of pending orders

Cancellations are different from the two previous types of orders studied (limit and market) because they are not a message to buy or sell some shares on the market, but a message to cancel a previous message to buy or sell some shares. For example, we cannot model the placement of cancellations as we did for the limit orders, since we can only cancel orders at prices where some orders are actually standing in the book. We thus adopt a completely different type of modelling for cancellations.

The first choice of modelling is that we do not model the intensity of submission of cancellation, but we model instead the lifetime of pending limit orders. One reason for this choice is that cancellations ensures the stability of the system. Cancellation process is intimately linked to the limit submission process. By defining an autonomous state-dependent cancellation process, we introduce a risk of instability in the model. The choice of the lifetime of orders as the main variable is thus a safe choice. Its drawback however is that it is a very difficult parameter to estimate. Our trades and quotes database does not provide a unique identifier for each order, thus when we observe a cancellation we do no know for sure which limit order has been cancelled. We can narrow it down by selecting only limit orders with the volume and price equal to the one cancelled, but this identification does not necessarily return a unique match. Finally, even if we perform the above algorithm with some selection rules, the obtained distribution is not necessarily easy to characterize. As an example, on the stock AIRP.PA on January 17th, 2011, the above algorithm gives an empirical distribution of lifetimes with median of 5.2 seconds, and a mean of 89.7 seconds.

We choose to compute the average lifetime of an order so that a basic order book model with Poisson intensities would have an average total liquidity in the book equal to the empirical observation. More precisely, Muni Toke (2015) shows that in an order book with Poisson arrival of market orders with intensity $\lambda^M \in \mathbb{R}_+$ and average size σ^M, Poisson arrival of limit orders with intensity $\lambda^L \in \mathbb{R}_+$ and average size σ^L, and a lifetime of pending limit orders exponentially distributed with parameter θ^{-1}, the expected total liquidity Q available in the book is

$$Q = \sigma^M \left(\frac{\nu}{q} - \delta + \frac{\delta q^{\frac{\nu}{1-q}}}{{}_2F_1\left(\delta, -\frac{\nu}{1-q}, 1+\delta, 1-q\right)} \right), \tag{14}$$

where $\nu = \frac{\lambda^L}{\theta}$, $\delta = \frac{\lambda^M}{\theta}$, $q = \frac{\sigma^M}{\sigma^L}$ and ${}_2F_1$ is the hypergeometric function. It is easy to numerically optimize θ so that Q given in the equation above is equal to its empirical counterpart.

The second choice of modelling deals with the "placement" of the cancellations. Mike & Farmer (2008) has proposed a three-variable model to determine the placement of cancellations, based on the distance to the best quote, the total liquidity available and the imbalance. We here propose a new efficient one-parameter model to choose which pending order is to be cancelled. We introduce as modelling variable the "priority index". We firstly define the "priority volume" of a limit order as the sum of all the sizes of pending limit orders standing ahead in the queue, i.e. at a better price

or at the same price but with time priority. If a limit order is the oldest order standing at the best quote, then it will be executed first when a market order arrives, its priority volume is thus zero. One may expect that the probability to be cancelled decreases with the priority volume, but that would be ignoring the fact that most of the activity occurs around the best quotes.

Let us now define the "priority index" ξ of a pending limit order as the ratio of the "priority volume" defined above over the total volume available in the book (on the same side). Obviously $\xi \in [0, 1]$. ξ can be used as a indicator of placement of cancellations inside the book. As for the empirical estimation of ξ however, our data does not allow for the unique tracking of individual orders. We know the price of an order, but not exactly where the order is inside the sub-queue of all orders at this price (at least not without further algorithmic development). We thus compute the priority volume as the total liquidity available at better prices plus half the liquidity available at the same price, i.e. we act as if the cancelled order were in the middle of the queue. This allows for an easy estimation of ξ on our data. It turns out that the distribution of cancellations as a function of ξ is remarkably smooth. Some empirical results are given below. We propose to model it with a scaled truncated power law distribution, i.e. we have the following model for the density of the cancellation "placement" $\pi^C : [0, 1] \rightarrow \mathbb{R}_+$:

$$\pi^C(\xi) = \frac{\sigma(\alpha + 1)}{(1 + \sigma)^{\alpha+1} - 1}(1 + \sigma\xi)^\alpha. \tag{15}$$

The log-likelihood of a sample $(\xi_1, \ldots, \xi_N)$ is straightforwardly computed as

$$\mathcal{L}(\alpha, \sigma) = N \log \left(\frac{\sigma(\alpha + 1)}{(1 + \sigma)^{\alpha+1} - 1} \right) + \alpha \sum_{i=1}^{N} \log(1 + \sigma\xi_i), \tag{16}$$

which can be numerically maximized using the `mle2` routine of the `bbmle` package. Numerical results of the maximum-likelihood estimation are given in Table 5. Illustrations of the quality of

RIC	α	(std)	σ	(std)
AIRP.PA	-1.378	*0.008*	6.760	*0.095*
ALSO.PA	-0.876	*0.005*	13.101	*0.241*
BNPP.PA	-1.256	*0.004*	16.014	*0.132*
BOUY.PA	-1.561	*0.017*	4.412	*0.098*
CARR.PA	-1.775	*0.015*	4.684	*0.078*
EDF.PA	-1.694	*0.013*	5.749	*0.090*

Table 5: Parameters for the placement of cancellations obtained by numerical minimization of the loglikelihood.

the fit are provided on Figure 11. Table and figures all show an excellent agreement between the model and the empirical data for all the stocks studied. Two graphs are shown for brevity. All
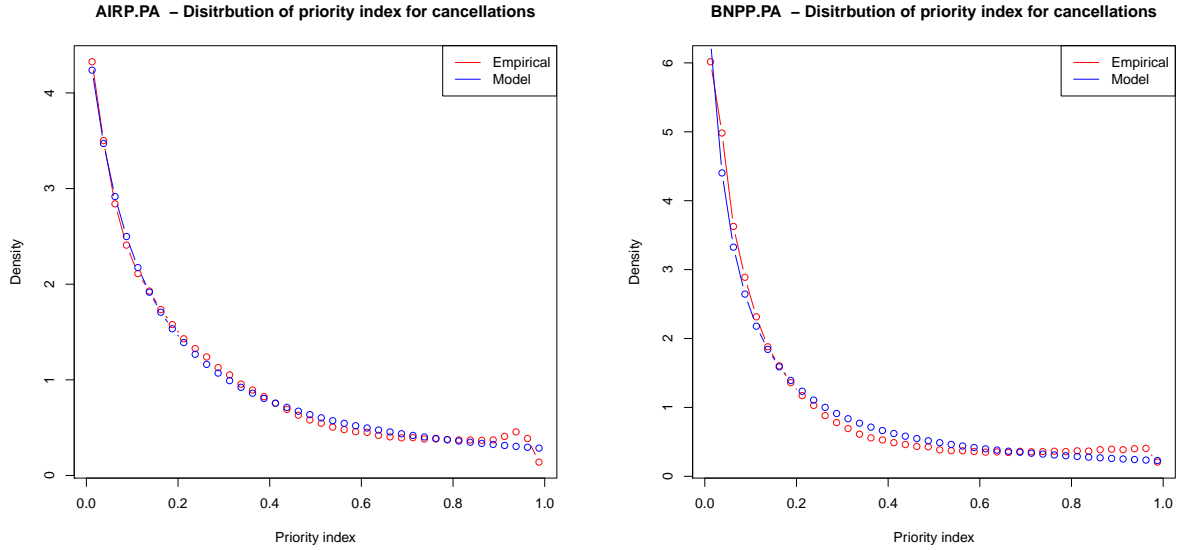
Figure 11: Empirical and model distribution of the placement of cancellations as a function of the priority index.

results not shown (ALSO.PA, BOUY.PA, CARR.PA, EDF.PA) are similar to AIRP.PA (left panel). BNPP.PA (right panel) is the worst visual fit of all 6 stocks tested (see also Section 6).

6 Stability of the fitted parameters across time

This work is primarily an exploratory investigation on the potential of a parametric model to reproduce the dynamics of an order book. Before going on with the simulation results of our model in Section 7, we report for the sake of completeness the results of the weekly calibration of the model on two and a half years of data, from January 2011 to June 2013, representing 128 trading weeks (two trading weeks with only three full trading days or less have been discarded) for the stock ACCP.PA (Accor Group, hotel industry, component of the CAC40). All results have been obtained after three numerical log-likelihood maximization with different starting points to avoid local extrema.

Results for the parameters of the intensities of market orders and limit orders are reported on Figure 12. These results confirms the exploratory work of Sections 3 and 4. All parameters are stable on the whole sample. Second-order parameters for both intensities lie in the range $[-1, 1]$, and first order parameters in the range $[-10, 10]$. Fitted parameters for market orders intensities are very stable across time. Confidence intervals are very thin for all parameters and therefore not reported for clarity. For market orders, β_1, β_{11}, β_2 and β_{12} are very stable and keep a constant sign through the sample, with a few exceptions. β_{22} is mostly positive until the first quarter of 2012,
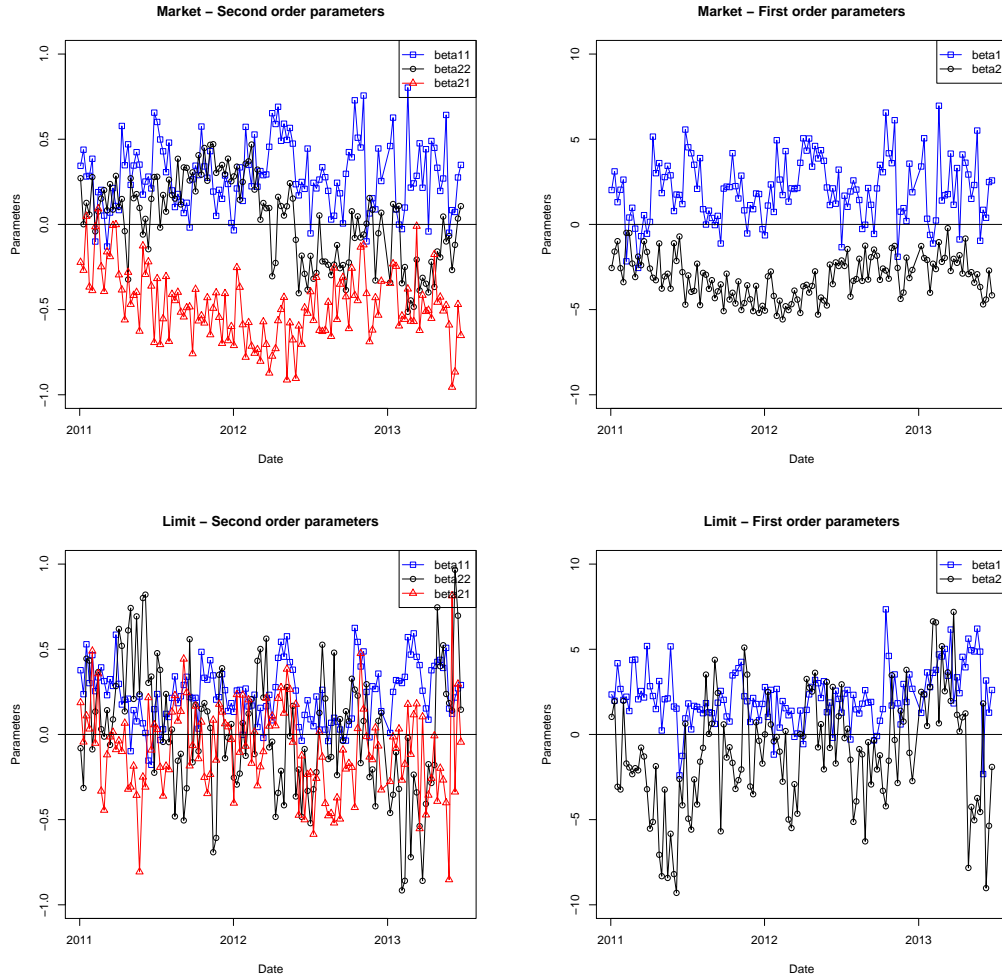
21

Figure 12: Second-order (left panels) and first-order (right panels) fitted parameters for the intensities of market orders (top panels) and limit orders (bottom panels) across time.
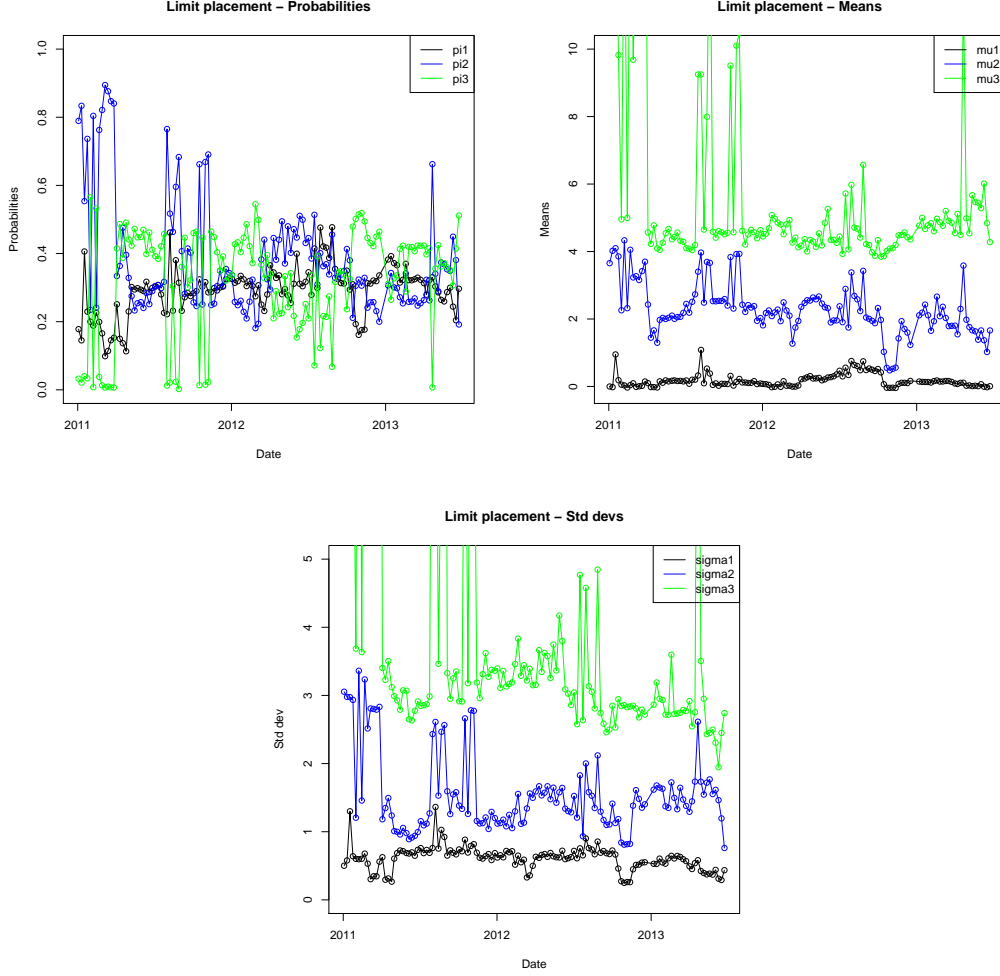
Figure 13: Fitted parameters of the Gaussian mixture model for the placement of limit orders across time.

and but mostly negative and more volatile afterwards. For limit orders, the same observations apply for the spread dependency, β_1 and β_{11}, being stable and keeping a constant sign across the whole two-year sample. Situation is less clear for the Q_{10} dependency, for which parameters are more volatile. All in all, this confirms the quality of the modelling of the spread dependency, and a greater variability for the dependence on the quantities, confirming the observations of Section 3 and 4.

Figure 13 plots the evolution across time of the fitted parameters for the placement of limit orders with the Gaussian mixture model described in Section 4.2. The three components of the placement distribution identified in Section 4.2 are clearly identified and stable across the two-year-and-a-half sample : a thin Gaussian at the best quote (in black), a larger one roughly two ticks
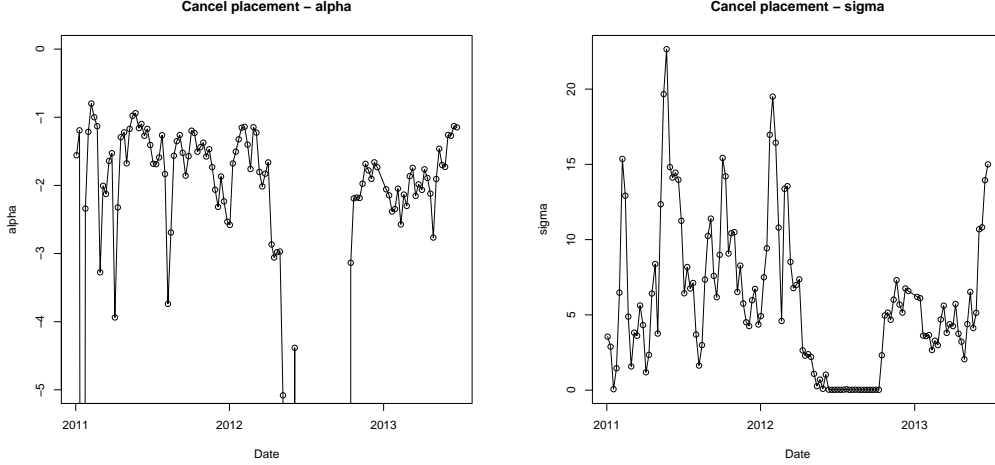
Figure 14: Fitted parameters of the cancellation model for the placement of limit orders.

inside the book, and a even larger one roughly five ticks inside the book. The component describing the placement of limit orders at the best quote is very stable, accounting for roughly 30% of the total distribution. For about one third of the weeks in 2011, mostly in the first quarter, it appears that two distributions seem sufficient to describe the placement of limit orders inside the book : in these weeks, the third component is large, far away in the book, and accounting for a very small part of the submitted limit orders. However on most of the sample, all components have similar weights accounting each for roughly one third of the orders.

Finally, the evolution of the fitted parameters for the cancellations are plotted in figure 14. It turns out that the exponent parameter α is quite stable across time, taking values in the interval $[-3, -1]$, except for four months in mid-2012 (June to September) during which the likelihood maximization gives a negative exponent two orders of magnitude below, associated with a scale parameter σ close to zero. Visual examinations of the fitting shows that in this period, the empirical distribution of the priority index of cancelled orders has a quite large drop for very small abscissas, i.e. that the intensity of cancellation at or close to the best quote is lower. As a consequence, the placement distribution has a maximum away from zero. In this context, it is natural that the performance of our single-exponent monotonous model distribution is degraded. This observation opens future directions to improve this model.

7   A market simulator with state-dependent order flows

Finally, this last section shows the benefits of our model by fitting it to daily empirical data and simulating it. Simulating a "realistic" limit order book is a quite complex task given the many parameters involved and the somewhat complex time-priority execution mechanism to be

24

implemented. Several results have previously been obtained, for example in Gatheral & Oomen (2010), Muni Toke (2011). Some key elements for basic simulation can be found in Abergel et al. (2016).

7.1 Market simulator

We build a market simulator with four agents. Two "liquidity providers" submit (and cancel) limit orders, one on the ask side and another on the bid side. Two "liquidity takers" submit market orders, one on the ask side and another on the bid side. We choose to simulate here a symmetric limit order book, i.e. both providers share the same parameters, and both takers share the same parameters.

Liquidity providers submit limit orders with the intensity $\lambda^L(S, Q_{10})$ defined in Equation (10). The distribution of the sizes of the limit orders is exponentially distributed with parameters $\frac{1}{\hat{\sigma}^L}$ where $\hat{\sigma}^L$ is the median of the empirical sizes of limit orders. The distribution of the prices of the limit orders is defined by our Gaussian mixture model given by Equation (12).

Liquidity takers submit market orders with the intensity $\lambda^M(S, q_1)$ defined in Equation (5). The distribution of the sizes of the limit orders is exponentially distributed with parameters $\frac{1}{\hat{\sigma}^M}$ where $\hat{\sigma}^M$ is the median of the empirical sizes of market orders.

Finally, cancellations in the order book occur with an intensity proportional to the available liquidity, i.e. $\lambda^C = Q\theta$ where Q is the total number of orders and θ is determined by the procedure detailed in Section 5 and Equation (14). When a cancellation occurs, a random priority index $\bar{\xi}$ is drawn according to the distribution with density π^C given at Equation (15). This distribution is easy to simulate given its inverse cumulative distribution function $(\Pi^C)^{-1}$:

$$(\Pi^C)^{-1}(x) = \frac{1}{\sigma} \left[\left[\left((1+\sigma)^{\alpha+1} - 1 \right) x + 1 \right]^{\frac{1}{\alpha+1}} - 1 \right]. \tag{17}$$

The order cancelled is then the first one that has a priority index greater or equal to $\bar{\xi}$.

7.2 Poisson simulator reference

To provide a reference simulation, we simulate a standard Poisson model. This reference model has the same agents, the same distributions of sizes of limit and market orders, and the same cancellation intensity proportional to the liquidity available. However, all agents submit their orders according to a homogeneous Poisson process with a constant intensity fitted by MLE estimation. The placement of limit orders is done according to the location-scale Student distribution given in Equation (11). Finally, the cancellation is purely zero-intelligence in the sense that the chosen order when a cancellation occurs is uniformly selected in the book.
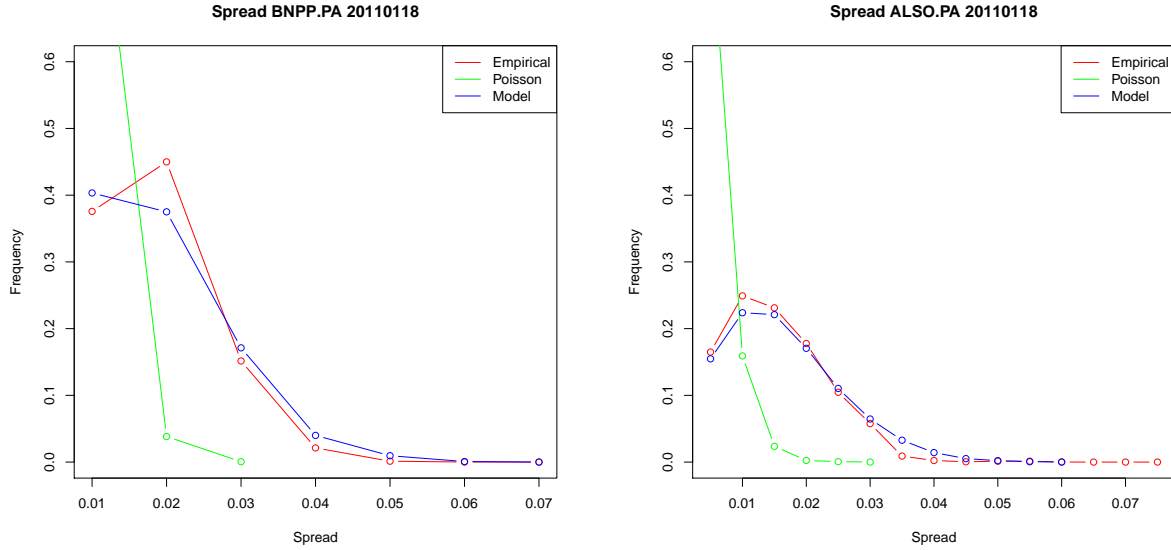
Figure 15: Distribution of the spread in the model, compared to the empirical distribution and the one produced by a Poisson model. Data: January 18th, 20011.

7.3   Simulation results

We fit our model for each stock of our sample, and using one day of trading. Since we simulate a symmetric limit order book, we aggregate bid and ask order flows in one sample for the fitting. We have made the full simulation of our model for each of the first two days of the sample, but for the sake of brevity, we show in this section the results for only one day, January 18th, 2011. Results for the other day tested are exactly similar. The sample used for fitting is smaller than the full one (ten days) used in the previous sections to derive the functional shapes of the intensities and distributions of our model. This may lead to potentially noisier estimates of our model, but for practical purposes one trading day is a convenient unit of time, hence this choice.

The simulator (and the reference Poisson simulator) is then run to produce exactly one day of trading data (i.e. the same length as the fitting sample). We then analyse the simulated data and compare it to the empirical observations.

One of the most important feature is that our model is able to reproduce very well the empirical distribution of the spread. Figure 15 shows two representative stocks, but all stocks tested show good fittings for the simulated distribution of our model, while the Poisson reference is not relevant at all. The spread in the Poisson model is most of the time equal to 1 tick, and the support of the whole Poisson simulated distribution is mostly $[1, 4]$ ticks, rarely larger, i.e. the book is "stuck". One can understand this result by observing that most of the limit orders are placed at or close to the best quote. When market orders are submitted (roughly 1 market order for 40 limit and cancel
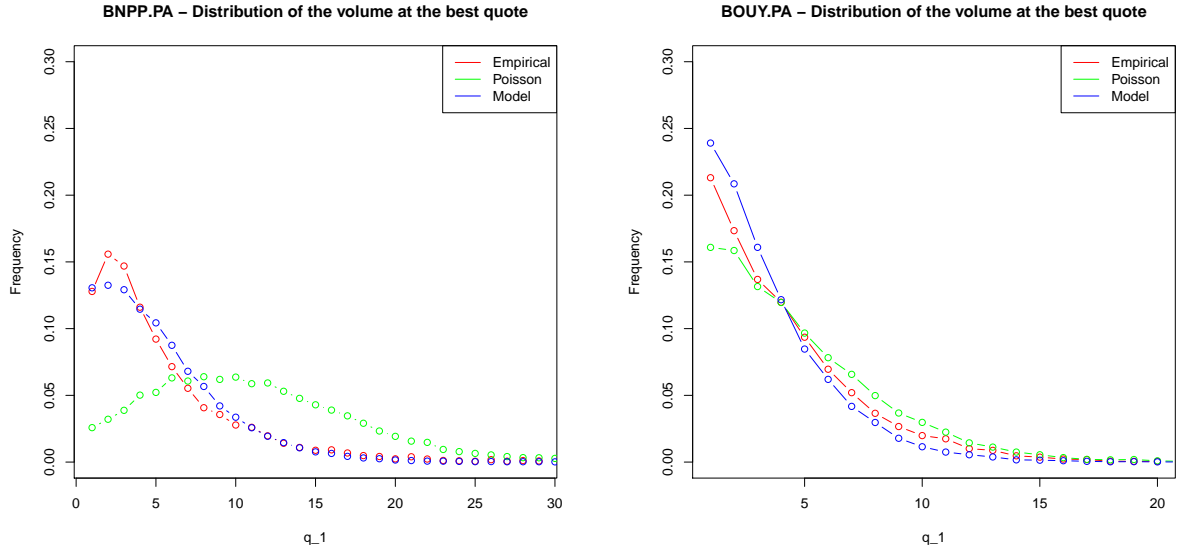
Figure 16: Distribution of q_1 in the model, compared to the empirical distribution and the one produced by a Poisson model. Data: January 18th, 20011.

orders), they are executed against a quite large volume available at the best quote, leaving the spread equal to one tick. In the Poisson model, there is no mechanism to cluster the market orders (see e.g. Muni Toke (2011)) and dig deeper into the order book, and hence get the limit order book out of this "small spread - high volume" state. On the opposite, our model of intensities is able to tackle this problem by increasing the market intensity and decreasing the limit intensity when the spread is small, as it is empirically observed in the previous sections. It also increases the intensity of submission of market orders when q_1 is small. Both mechanisms increase the probability to observe states of the limit order book with larger spread. It is remarkable to observe that this close fit is obtained for all stocks and dates tested, irrespective of the liquidity and ticksize of the stock studied.

We now turn to the second modelling variable of our model, q_1, which is closely linked to the spread. On Figure 16, we plot the empirical distribution of q_1 and its simulated counterparts. There again, the model provides an excellent fit for this distribution while the standard Poisson reference constantly underestimates the probability to observe smaller values of q_1, i.e. its q_1 distribution is shifted to the right. Two representative stocks are represented: AIRP.PA, ALSO.PA and CARR.PA (not shown) are similar to BNPP.PA (left panel). The best observed result for the Poisson model is for BOUY.PA (right panel), still worse than the state-dependent model. The failure of the Poisson model is linked to the above analysis of the spread results. In the Poisson model, since the spread is stuck to 1 or 2 ticks, most of the limit orders are submitted at or around the non-moving best quotes, hence increasing the available volume at the best quote, which we now observe on Figure
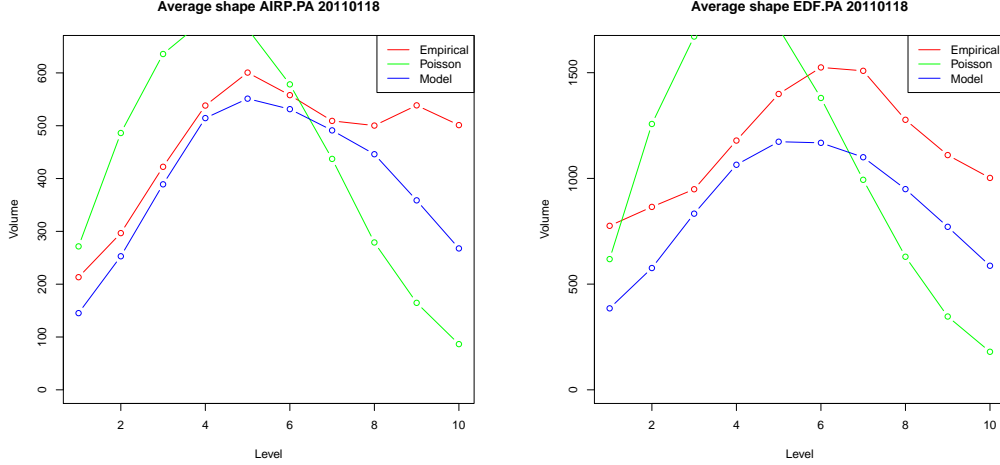
27

Figure 17: Average shape of the order book in the model, compared to the empirical shape and the one produced by the Poisson model. Data: January 18th, 2011.

16. This will in turn prevent market orders from increasing the spread, hence a vicious circle for the Poisson model. This analysis shows that the modelling of the spread and the first limit are closely linked problems in our limit order book framework.

We finally turn to the third variable used in our model, the total volume available Q_{10}. We start with the average order book shape, i.e. the time average of the quantities q_1 to q_{10}, the sum of which is Q_{10}. Figure 17 plots the empirical average shape of the order book and the ones produced by the simulators for two representative stocks, but all results are similar. Both models are able to quite well reproduce the order of magnitude of average shape of the limit order book. This is not surprising, since the magnitude of the average is directly linked to the way we estimate the parameter θ in Section 5, which is identical in both models. However only our model correctly reproduces the slope of the average order book for the best prices, as well as a sound estimation of the position of the maximum away from the best quotes: BNPP.PA, BOUY.PA, CARR.PA and ALSO.PA (not shown) are roughly similar to BNPP.PA (left panel), while EDF.PA (right panel) is the worst visual fit of our model. As for the Poisson reference, it exhibits a sharper slope for the best prices, realizes a maximum too high and too close to the best quote, and underestimates the volume available far away from the best quotes. Once again, these observations are valid for all stocks and dates tested.

Beyond the analysis of these averages values, Figure 18 plots the empirical distribution of Q_{10} for two representative stocks (again all results are similar). It turns out that if the model is able to reproduce the order of magnitude of the mean of Q_{10}, the support of the distribution is actually smaller and slightly right-shifted compared to the empirical distribution. In fact the empirical Q_{10} distribution exhibits a quite heavy tail for large values of Q_{10}. Since both models are fitted on the
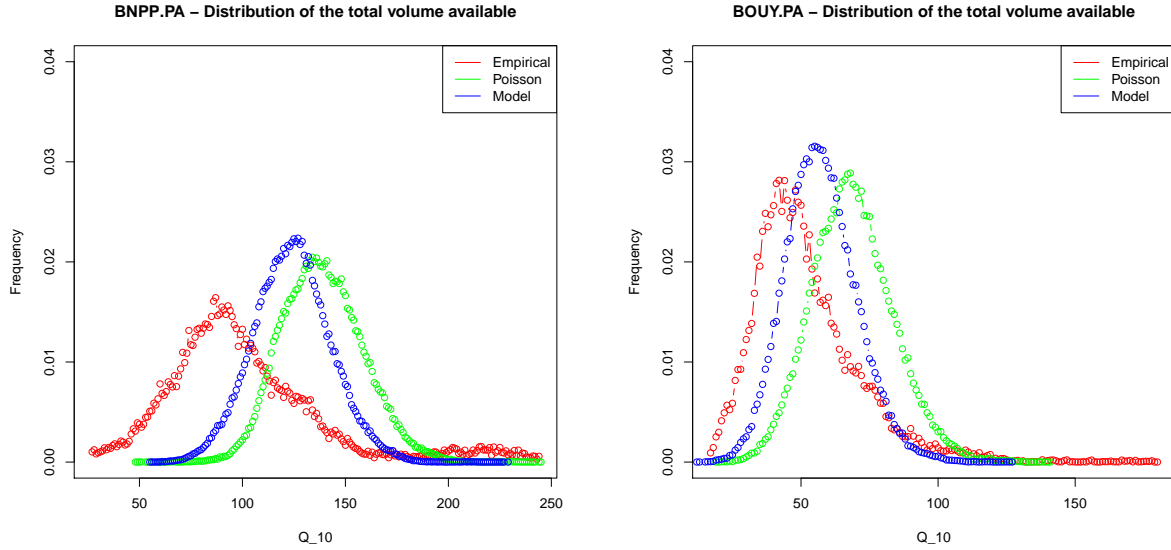
28

Figure 18: Distribution of Q_{10} in the model, compared to the empirical shape and the one produced by the Poisson model.

mean, this leads to an underestimation of the probability of lower values of Q_{10} in both simulations. Note however that the full model outperforms the Poisson reference even in this case. Figure 18 suggests that some state-dependency of the intensities is still not grasped by the model, and that there is probably room for more complex definitions of theses intensities. Compared to the Poisson reference, our model is able to reproduce a regime with lower liquidity in the book, by increasing the intensity of limit orders when the liquidity becomes low, decreasing it when the book gets full. Figure 18 suggests that this phenomenon is even more pronounced in reality : most of the time the liquidity available in the book is even lower than in our model, with rare occurences of very large volumes, impacting the mean.

8   Conclusion

We have provided a fully parametric model for the limit order book. The submission of orders is modelled as a point processes with state-dependent intensities. We provide detailed functional forms for these intensities, as well as the estimation procedure by likelihood maximization. By developing a market simulator we are able to show that the model performs very well to reproduce key features of the order book, such as the spread, the volume of the best quote, and the average shape of the order book.

This very empirical and numerical work will hopefully lead to further improvements. The intensities we have proposed here are chosen with respect to some model principles in the choice of

variables and functional forms. One may probably go further in the statistical model by experimenting other forms or variables. In particular, the model for the placement of cancellations is monotonous in its main variable, which is sufficient on the primary sample tested, but leaves room for further work and improvements on more recent data.

This work could also stimulate research on the stability of such complex random systems. Although the mathematics of the "Poisson" models for the order book are beginning to be well-understood, the introduction of state-dependent intensities could lead to several theoretical problems that have not been studied here.

Acknowledgements

References

Abergel, F., Anane, M., Chakraborti, A., Jedidi, A. & Muni Toke, I. (2016), *Limit order books*, Cambridge University Press.

Abergel, F. & Jedidi, A. (2013), 'A mathematical approach to order book modeling', *International Journal of Theoretical and Applied Finance* **16**(05), 1350025.

Challet, D. & Stinchcombe, R. (2001), 'Analyzing and modeling 1+ 1d markets', *Physica A: Statistical Mechanics and its Applications* **300**(1), 285–299.

Cont, R. & De Larrard, A. (2012), 'Order book dynamics in liquid markets: limit theorems and diffusion approximations', *arXiv preprint arXiv:1202.6412* .

Cont, R., Stoikov, S. & Talreja, R. (2010), 'A stochastic model for order book dynamics', *Operations research* **58**(3), 549–563.

Gatheral, J. & Oomen, R. C. (2010), 'Zero-intelligence realized variance estimation', *Finance and Stochastics* **14**(2), 249–283.

Huang, W., Lehalle, C.-A. & Rosenbaum, M. (2015), 'Simulating and analyzing order book data: The queue-reactive model', *Journal of the American Statistical Association* **110**(509), 107–122.

Huang, W. & Rosenbaum, M. (2015), 'Ergodicity and diffusivity of markovian order book models: a general framework', *arXiv preprint arXiv:1505.04936* .

Mike, S. & Farmer, J. D. (2008), 'An empirical behavioral model of liquidity and volatility', *Journal of Economic Dynamics and Control* **32**(1), 200–234.

Muni Toke, I. (2011), "Market making" in an order book model and its impact on the spread, *in* 'Econophysics of Order-driven Markets', Springer, pp. 49–64.

Muni Toke, I. (2015), 'The order book as a queueing system: average depth and influence of the size of limit orders', *Quantitative Finance* **15**(5), 795–808.

Muni Toke, I. (2016), 'Reconstruction of order flows using aggregated data', *to appear in Market Microstructure and Liquidity* .

Smith, E., Farmer, J. D., Gillemot, L., Krishnamurthy, S. et al. (2003), 'Statistical theory of the continuous double auction', *Quantitative finance* **3**(6), 481–514.