Clustering of order arrivals, price impact and trade path optimisation

Patrick Hewlett*

May 6, 2006

Abstract

We fit a bivariate Hawkes process to arrival data for buy and sell trades in FX markets. The model can be used to predict future imbalance of buy and sell trades conditional on history of recent trade arrivals. We derive formulae for the raw price impact of a trade as a function of time assuming that trade arrivals are governed by a Hawkes process and that the price is a martingale, and show that the price impact of a series of trades is given by superposition of their individual price impacts. We use these formulae to parameterise a model for optimal liquidation strategies.

1 Introduction

It is a well known feature of many financial markets that trading activity tends to cluster in time, and that trades of the same sign tend to cluster together in the sequence of buys and sells. Such clustering can be modelled with a multivariate point process. Liquidity providers[1] ("market makers") in the FX market are well aware of this clustering, and anecdotal evidence suggests that they pay close attention to the pattern of arrivals of buy and sell orders when setting prices.

In this paper we focus on a model in which order arrivals are governed by a special class of point process, the Hawkes self-exciting process; in this case, the mathematical solution of the market maker's problem has a particularly tractable form. We also see that the Hawkes process can be fitted quite successfully to empirical order arrivals data.

In contrast to liquidity providers, liquidity demanders ("traders") often split large trades into multiple tranches over a period of minutes or hours, in order to alleviate market impact costs. One of their concerns is that market makers will

*OCIAM, 24-27 St Giles, Oxford OX1 3LB; tel 01865 276502; email: hewlett@maths.ox.ac.uk. Thanks are due to Sam Howison and Giovanni Pilliterri for helpful comments on this version, and to Clive Bowsher for advice on Hawkes processes. Financial support from Deutsche Bank and the EPSRC is gratefully acknowledged.

[1] The FX market is arranged as an electronic limit order book, so all participants have the opportunity to act as market makers. In what follows we assume that there is a single market maker who sets competitive prices.

guess from their pattern of trading that they are in the process of executing a large trade, and penalise them accordingly through less favourable prices. This activity is formulated in our model as the market makers trying to predict future trading from past trading. We assume that market makers form their expectations of the timing and direction of future trading based on a Hawkes process model, and that they set prices competitively based on these expectations.

The trader is then faced with the usual dilemma: trade too quickly and suffer severe market impact costs; trade too slowly and run the risk of adverse price movements before the trade is completed. Whilst models describing this dilemma are well developed, methods of parameterising them are not. This work offers a method of parameterisation which only requires data on the trade arrival process, which is readily available to FX market participants.

It should be emphasised that we do not attempt to offer a game-theoretic solution for the behaviour of the trader and market-maker. Instead, we take the empirical clustering of trades as given, compute how the market-maker ought to react to trading patterns assuming that the clustering is well described by a Hawkes process, and finally compute how the trader ought to behave given the market-maker's (empirically) rational response to trading. This allows us to build a practically applicable model of the trading environment without needing to model the heterogeneity of traders and their motivations.

This paper is organised as follows. Section 2 reviews past work on trade arrival processes and multiple-tranche trading optimisation. In Section 3 we present the model for the market-marker's problem and discuss the mathematical properties of Hawkes processes that allow a solution to be developed. In Section 4 we describe the dataset and fit the Hawkes model. In Section 5 we describe the trader's problem and compute solutions using the parameters obtained above. Section 6 concludes.

2 Literature Review

An introduction to the mathematical theory of point processes is given in [8]. Hawkes processes, first introduced by [23], are a particularly tractable type of point processes for our purposes, because closed form expressions exist for the expected future number of arrivals of each type and current intensity given the observed history of the process. Estimation is also relatively straightforward (see e.g. [28]). The state of the art in estimation and model validation for multivariate Hawkes processes in finance is summarised in [6].

There is a growing literature on the application of point processes to high frequency financial data. Important early work focuses on modelling inter-event duration ([13], [14]) and the effect of duration on price impact and trade sign autocorrelation ([10]). [12] extend these models to allow for the censoring of quotes after a trade by intervening trades. We are concerned here with predicting future trades given the pattern of past trades. These duration-based models are somewhat intractable for this purpose in a multivariate context.

In high frequency finance, rounding of times to the nearest second is common. A half-way house between true point process modelling and discrete time modelling is obtained by binning data into discrete intervals (which may or may not correspond to the measurement frequency). [22] and [9] approach modelling of trades and quotes in this spirit. Whilst we believe that we could have developed our model along these lines, we felt that the advantages of using "true" continuous time processes outweighed possible disadvantages[2].

The use of Hawkes processes in particular for modelling order arrivals and microstructure scale price movements is investigated by [7], [6], [25], [3]. Elsewhere in the mathematical finance literature, Hawkes processes have been advocated for modelling of credit contagion [18] and clustering of extreme price moves [26].

Much thinking on the optimal way to split a large order into tranches over time is based on ideas advanced in [4], [2]. Despite (or perhaps because of) the practical relevance of such models, there is little published on how to parameterise them. [1] details one strategy for parametrisation, which requires as data the ex-post cost of various trading strategies. By contrast, the model we present in Section 5 could in principle be parameterised using only data on the times and directions of order arrivals.

3 A self-exciting model of trade arrivals

Intensity-based approaches focus on arrival intensities for the counting processes $N_t^{(i)}$. Arrival intensity $\lambda_t^{(i)}$ conditional on a filtration $\mathcal{F}_t$ is defined by

$$\lambda_t^{(i)}|\mathcal{F}_t = \lim_{\delta t \to 0} \frac{1}{\delta t} \mathbb{E}(N_{t+\delta t}^{(i)} - N_t^{(i)}|\mathcal{F}_t)\,. \tag{1}$$

In the case of purely self-exciting processes, the intensity is a functional of past arrivals[3]. For a linear self-exciting process, we have

$$\lambda_t^{(i)} := \mu^{(i)} + \sum_{j=1}^{I} \int_{u<t} h_{ij}(t-u)\, dN_u^{(j)}\,. \tag{2}$$

Here μ_i can be understood as the "base" intensity of arrivals of type i (the intensity if there have been no past arrivals of any type), and h_{ij} the propagator of an arrival of type j onto the intensity of arrivals of type i in the future. We will be interested in parametric forms for g — in particular we consider the case where g is a sum of exponentials:

$$h_{ij}(s) = \sum_{k=1}^{K} \alpha_{ij} e^{-\beta_{ij} s}\,, \tag{3}$$

so that

$$\lambda_t^{(i)} = \mu^i + \sum_{k=1}^{K} \sum_{j=1}^{I} \int_{u<t} \alpha_{ij} e^{-\beta_{ij}(t-u)}\, dN_u^{(j)}\,. \tag{4}$$

This specification is labelled a Hawkes-E(K) process in [6]. Other forms for h have been advocated in the finance literature (e.g. power law, Laguerre polynomial). The advantage of the exponential specification is that the likelihood function can be

[2] In particular, there is well-developed mathematical machinery for continuous time point processes, whereas the models of [9] ("CBin models") are less developed; whilst the mathematical foundations of point processes may appear to present a barrier to practitioners, the relevant properties can safely be understood intuitively in this context. Discrete sampling does create a few problems in the busiest FX markets (e.g. USD/EUR) where the probability of several events in one second is significant.

[3] in contrast to doubly stochastic processes where there is an unobserved component driving the intensity.

computed in $O(N)$ steps, whereas for more general g, $O(N^2)$ steps will be required. The importance of long-range dependence and the extent to which a mixture of a small number of exponentials provides a satisfactory parameterisation is an area for future research.

3.1 Time domain properties

We now review the time domain properties of a Hawkes E(K) point process that we will require to develop our model of price formation. We assume that the conditions for stationarity (see e.g. [8]) are satisfied.

3.1.1 Expected arrivals following a single arrival

When a new arrival occurs, expectations of future arrivals are increased. The new arrival is expected to cause future arrivals directly, via h; these are then expected to cause further arrivals, so that the original arrival can be thought of as causing a cluster or cascade ([24]). Let $f_{ij}(s)$ denote the increase in intensity of arrivals of type i caused (directly and indirectly) by an arrival of type j at time zero. Then f satisfies the following integral equation:

$$f_{ij}(s) = h_{ij}(s) + \sum_{l=1}^{I} \int_0^s h_{lj}(u) f_{il}(s-u)\, du\,. \tag{5}$$

It is easily checked that in the univariate case with a single decay timescale $I = 1, K = 1, h(s) = \alpha e^{-\beta s}$, a solution is

$$f(s) = \alpha e^{-(\beta-\alpha)s}\,. \tag{6}$$

In the multivariate case and and the case where there are multiple decay timescales, closed form solutions are still available [23]. If h does not have exponential form, closed form solutions are not necessarily available but direct numerical solution is straightforward.

3.1.2 Prediction of arrival rates given history

We now consider the problem of predicting the total number of arrivals of each type arising in some (t, T) given information at time t. Writing $\kappa^{(i)}(s) = \mathbb{E}(\lambda^{(i)}(s)|\mathcal{F}_t)$, we have:

$$\kappa^{(i)}(s) = \mu^{(i)} + \sum_{j=1}^{I} \int_{v<t} h_{ij}(s-v) dN_v^{(j)} + \sum_{l=1}^{I} \int_t^s \kappa^{(l)}(u) f_{il}(s-u)\, du\,. \tag{7}$$

We exploit the linearity of (7) to search for a solution of the form

$$\kappa(s) = \kappa_0(s-t) + \int_{v<t} G(s-t; t-v)\, dN_v\,, \tag{8}$$

where

$$\kappa_0(x) = \mu + \int_0^x \kappa_0(\xi) f(x-\xi)\, d\xi \tag{9}$$

$$G(x, y) = h(y+x) + \int_0^x G(\xi; y) f(x-\xi)\, d\xi\,. \tag{10}$$

We will use this representation to compute the theoretical price impact function for a given order arrival process. Only for special types of point process can we write the conditional intensity given observed information in the convenient form (8). For example, for the autoregressive conditional duration model of [13], such a representation is not possible.

In the univariate single timescale case, we note that (10) and (5) differ only by a factor of $e^{-\beta y}$ in the first term on the right hand side. So in this case we have $G(x;y){=}e^{-\beta y}f(x)$. In the multivariate and multiple timescale case, we suppose that closed form solutions for κ_0 and G would still be computable using Laplace transform methods (Cf. [23]), although practitioners may be just as comfortable with a numerical solution, which in any event only needs to be computed once. In the non-exponential case, numerical solution is usually mandatory.

3.1.3 Conditional arrival rates

The number of arrivals caused directly or indirectly by an arrival of given type is not directly observable in a sample time series of a self-exciting process, because arrivals are also correlated with past arrivals. In order to calculate a longitudinal conditional expectation, we need to use a two-sided version of (5). Let $c_{ij}(s)$ be the intensity of arrivals of type i at time $t+s$ conditional on an arrival of type j at time t, defined[4] for $s \neq 0$. Then

$$c_{ij}(s) = h_{ij}(s) + \sum_{l=1}^{I} \int_{-\infty}^{s} c_{lj}(u) h_{il}(s-u)\, du\,. \tag{11}$$

Again, in the single timescale univariate case a straightforward solution exists. We have

$$c(s) = \mu \frac{\beta}{\beta - \alpha} + \frac{2\beta - \alpha}{2(\beta - \alpha)} \alpha e^{-(\beta-\alpha)s}\,. \tag{12}$$

After scaling by the long-run mean arrival rate for the conditioning event, the function $c(s)$ is known as the covariance density of the process. It represents the joint arrival intensity of events separated in time by s. A sample covariance density can be computed using histogram methods.

3.2 Efficient price process

We now consider how market-makers would set prices if order arrivals were governed by a Hawkes process. We denote the counting processes for arrival of buy and sell orders by N_t^{buy}, N_t^{buy} respectively. We suppose that the order arrival process $(N_t^{\text{buy}}, N_t^{\text{sell}})$ is a bivariate Hawkes process, and denote cumulative order imbalance by $F_t := N_t^{\text{buy}} - N_t^{\text{sell}}$. The functions κ_{buy}, G_{buybuy} etc. are all defined by indexing processes by $\{\text{buy}, \text{sell}\}$ in place of $\{1, ..., I\}$.

Our assumptions regarding order submission and price formation are as follows:

1. All individual market buy and sell orders are of a standard size, w.l.o.g. 1. Only one trade can take place at any instant.

[4] alternatively, we could add $\delta(s-0)$ to $c_{ii}(s)$ and subsume the first term in the equation into the integral

2. At the end of the trading period, which is assumed to be far away compared with the timescale of self-excitement of the Hawkes process, the asset is valued at some initial price P_0 plus θ times[5] the imbalance of buy orders over sell orders during the trading period.
3. Market makers are risk neutral, and act to maximise the sum of cash and inventory value at the end of the trading period.
4. Market makers are perfectly competitive.
5. Market makers set a single bid price, and a single ask price, good for one unit of asset, and are entitled to revise these prices after every trade.
6. The parameters governing the buy and sell processes are symmetric[6].

By Assumptions 3 and 4, market-makers set prices at time t based on their expectations of the price at the end of the trading period, based on observed order flow at time t. As in e.g. [20], the conditioning information for the bid and ask prices includes the direction of any trade at time t, whose size is known in advance to be unity by Assumption 1. Assumption 2 implies that the relevant expected terminal value is $\mathbb{E}(P_0 + \theta F_\infty)$. So

$$\begin{aligned} P_t^{\text{mid}} &= P_0 + \mathbb{E}(\theta F_\infty | \mathcal{F}_{t^-}) = P_0 + \theta F_t + \mathbb{E}(\theta(F_\infty - F_t)|\mathcal{F}_{t^-}) && (13)\\ P_t^{\text{bid}} &= P_0 + \mathbb{E}(\theta F_\infty | \mathcal{F}_{t^-}, \text{Market sell at t}) && (14)\\ P_t^{\text{ask}} &= P_0 + \mathbb{E}(\theta F_\infty | \mathcal{F}_{t^-}, \text{Market buy at t}) && (15) \end{aligned}$$

3.2.1 The price impact function

We now show (Cf [5]) that the mid price can be written in the form

$$P_t^{\text{mid}} = P_0 + \theta \int_0^t I(t-s)\, dF_s\,, \tag{16}$$

where $I(\cdot)$ is the impact of a trade through time. We would expect the initial impact of a trade to be large, corresponding to market-makers' (justifiable) fears that it is likely to be followed by more trades of the same sign. As time passes, the probability of further trades of the same sign decreases, and we would expect the impact to decay, eventually reaching the permanent impact θ.

To calculate I, we note first that

$$\mathbb{E}(F_\infty - F_t|\mathcal{F}_t) = \mathbb{E}(N_\infty^{\text{buy}} - N_\infty^{\text{sell}} - N_t^{\text{buy}} + N_t^{\text{sell}}|\mathcal{F}_t) \tag{17}$$

can be calculated by integrating out the difference between predicted buying and selling intensities:

$$\mathbb{E}(F_\infty - F_t|\mathcal{F}_t) = \int_s^\infty \kappa^{\text{buy}}(s) - \kappa^{\text{sell}}(s)\, ds \tag{18}$$

[5] θ is the permanent price impact of a market order. It can be estimated by regressing price movements of cumulative order flow with a fairly large sampling interval (e.g. one day), to remove distortions due to temporary price impact. It is sometimes known as Kyle's λ; in this paper λ is reserved for intensity.

[6] This, together with stationarity, guarantees that the expected order imbalance at ∞ exists. We could in principle build an asymmetric model, but would then have to impose a finite horizon, which is inconvenient. Such a model might be practically useful, for example where it is known in advance that there will be a bias towards selling all day.

We then apply (8), noting that the κ_0 terms for buying and selling cancel because we have assumed symmetry of parameters. We get:

$$\begin{aligned}\mathbb{E}(F_\infty - F_t|\mathcal{F}_t) = \int_s^\infty \int_0^t & G_{\text{buybuy}}(s; t-u) - G_{\text{buysell}}(s; t-u) dN_u^{\text{buy}} \\ &+ \int_0^t G_{\text{buybuy}}(s; t-u) - G_{\text{sellbuy}}(s; t-u) dN_u^{\text{sell}}\, ds\,,\end{aligned} \tag{19}$$

with G_{buybuy} etc. defined as in (10). Noting that the first term in (13) can be written

$$F_t = \int_0^t (dN_u^{\text{buy}} - dN_u^{\text{sell}})\,, \tag{20}$$

we get

$$I(x) = \theta\left(1 + \int_x^\infty G_{\text{buybuy}}(u) - G_{\text{buy}y\text{sell}}(u)\, du\right)\,. \tag{21}$$

In the case where buying and selling are governed by independent exponential self-exciting processes with parameters α and β we have

$$I(x) = \theta(1 + \frac{\alpha}{\beta - \alpha} e^{-\beta x}) \tag{22}$$

3.2.2 Bid and ask prices

Inspecting (14) and (15), we see that the bid and ask prices differ from the mid price only by the expected price movement that a trade arrival causes, that is by $I(0)$. We note that in this model, the information conveyed in a trade arrival does not depend on the past history of trades. Econometric models such as [10] have the property that the price information conveyed in a trade tends to be greater at times when trading is more active. Introduction of such a feature would require us to make the trade arrival process and price response nonlinear.

4 Estimation

We now fit a Hawkes process to trade arrival data recorded from Reuters D-2000 for the currency pair EUR/PLN. In order to get a handle on the kinds of effects we are trying to model, we first examine a histogram of empirical conditional intensity of buys and sells following a buy trade (Figure 1 - left). This histogram is effectively a non-parametric estimate of the function $c(\cdot)$ in (11). We see that following a buy there is a large (initially around threefold) increase in the intensity of buys, and a small increase in the intensity of sells. In our self-exciting model, this should correspond to a large self-exciting effect (large positive α_{buybuy}) and a small cross-exciting effect[7] (small positive α_{buysell}). Both effects appear to have a characteristic timescale (β^{-1}) of a few minutes.

[7] Since these intensities are not deseasonalized, the apparent cross-excitement could be partially due to seasonality effects — but the separation of timescales makes this unlikely

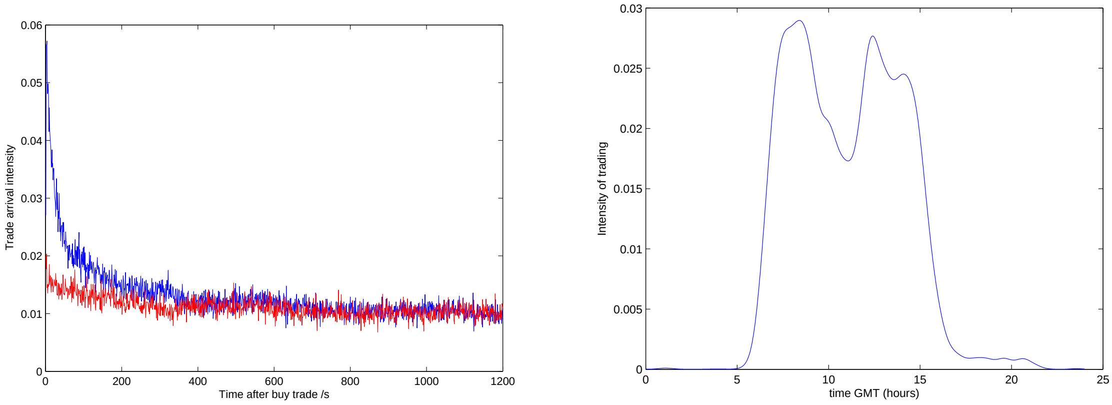

Figure 1: Empirical conditional intensity of buys and sells following a buy (left) estimated with a bin size of 1 second; daily seasonality (right) estimated by Kernel regression with bandwidth 1/2 hour.

4.1 Data and cleaning

The full dataset consists of records of 25,629 market orders recorded on EBS over two months for EUR/PLN. Each transaction is time-stamped to the nearest second, and marked “buy” or “sell” according to the direction of the market order giving rise to the transaction. Where a market order is submitted that requires clearing against limit orders at several different prices, it is recorded as several trades.

We have no record of the volume of trades. Market participants see the same raw sequence of buys and sells as we are using, without volume. They could in theory infer volumes of trades partially[8] by careful examination of changes in quoted volumes at best bid and ask in the limit order book. We hope to have the same limit order book data as market participants in the future; when this becomes available, we will incorporate volume into the model.

We aggregate simultaneous trades, so that our fitted model makes no distinction between a buy trade that eats up only the best ask and a buy trade that eats up several ticks of liquidity. An alternative procedure would have been to add 1 second to all inter-arrival times. This however would have introduced spurious predictability at short timescales.

Examining a Kernel regression of trading intensity by time of day (Figure 1 - right), we see something of a U-shaped pattern during the trading day, but note that the variation in intensity by time of day is less than a factor of two between 0700 and 1500. We therefore fit a time-homogeneous Hawkes model[9] to trades taking place between these times. Noting the separation of timescales between the self-excitement effect (which turns out to have a timescale of a few minutes) and variations in intensity due to time of day gives use some confidence that the time-homogeneous model is reasonable approximation at the proof of principle stage.

We then transform calendar time by gluing consecutive days together, so that

[8] Accurate volume figures would not be available because of the way that the limit order book is displayed on EBS dealing screens, and because of the difficulty in distinguishing market orders and cancellations.

[9] Time of day effects can be incorporated fairly easily.

1500 on Monday is followed immediately by 0700 on Tuesday in transformed time. Whilst this procedure will introduce unwanted edge effects, the loss of accuracy will be small, again because of the separation of timescales. See [6] for a more accurate model of inter-day effects — the methodology in that paper could easily be extended to our case.

4.2 Estimation

We now specialise (2) to choose the simplest model which will allow for self- and cross-excitement:

$$\lambda_t^{\text{buy}} = \mu_{\text{buy}} + \alpha_{\text{buybuy}} \int_{u<t} e^{-\beta_{\text{buybuy}}(t-u)}\, dN_t^{\text{buy}} + \alpha_{\text{buysell}} \int_{u<t} e^{-\beta_{\text{buysell}}(t-u)}\, dN_t^{\text{sell}} \tag{23}$$

$$\lambda_t^{\text{sell}} = \mu_{\text{buy}} + \alpha_{\text{sellsell}} \int_{u<t} e^{-\beta_{\text{sellsell}}(t-u)}\, dN_t^{\text{sell}} + \alpha_{\text{sellbuy}} \int_{u<t} e^{-\beta_{\text{sellbuy}}(t-u)}\, dN_t^{\text{buy}} \tag{24}$$

and impose symmetry constraints $\mu_{\text{buy}} = \mu_{\text{sell}} =: \mu$, $\alpha_{\text{buybuy}} = \alpha_{\text{sellsell}} =: \alpha_{\text{same}}$, $\alpha_{\text{sellbuy}} = \alpha_{\text{buysell}} =: \alpha_{\text{cross}}$, $\beta_{\text{buybuy}} = \beta_{\text{sellsell}} =: \beta_{\text{same}}$, $\beta_{\text{sellbuy}} = \beta_{\text{buysell}} =: \beta_{\text{cross}}$, so that we must estimate five parameters. We fit our model to the data to trades in the week[10] commencing 7th May 2005, aggregating simultaneous trades. The number of trades in the cleaned sample is 2,308.

We estimate our model parameters using maximum likelihood [28], subject to a non-negativity constraint[11] and find that $\mu = 0.0033s^{-1}$, $\alpha_{\text{same}} = 0.0169s^{-2}$, $\alpha_{\text{cross}} = 0$, $\beta_{\text{same}} = 0.0286s^{-1}$ and β_{cross} is not identified as $\alpha_{\text{cross}} = 0$. We also estimate a pure Poisson model, $\alpha_{\text{same}} = \alpha_{\text{cross}} = 0$, and find that $\mu = 0.0080$.

The maximised log-likelihood for the Hawkes model exceeds a pure Poisson log-likelihood by 1,017. The usual likelihood ratio test does not (quite) apply in this case due to the non-negativity constraints on the parameters. If it did, a likelihood ratio statistic of more than 7.54 would be sufficient to prefer the Hawkes model over a pure Poisson model at the 1% level. A likelihood ratio of 1,017 is sufficiently large that we (informally) prefer the Hawkes model to the null hypothesis of a Poisson model.

For intensity based point process models in general, the intensity-weighted waiting times $\int_{T_{n-1}}^{T_n} \lambda_{\text{buy}}(t) + \lambda_{\text{sell}}(t)\, dt$ between arrival times T_n of consecutive events have a standard exponential distribution[12]. Below we show QQ plots of these intensity-weighted waiting times for the fitted Hawkes model, and the Poisson model for comparison. We see that the fit for the Hawkes model is satisfactory, and that the fit for the Poisson model is poor, except perhaps in the upper tail. An explanation for the acceptable fit of the Poisson model in the upper tail is that larger waiting times tend to occur at times when the process is unexcited; at these times the behaviour of a self-exciting process is more Poissonian.

By way of illustration of behaviour of the fitted Hawkes model, the fitted intensities are shown over one-day and one-hour windows in Figure 3.

[10] We have used a reduced sample to save on computational time — we are not aware of any open-source software for computing MLEs for Hawkes processes and therefore have used our own MATLAB code. Numerical optimisation using MATLAB's *fmincon* function took ten minutes.

[11] In such a linear Hawkes model non-negativity of α is required to keep intensity positive at all times

[12] Another way of saying this is that under the stochastic time change $t \mapsto \int_0^t \lambda(\tau)d\tau$ the process becomes a standard Poisson process

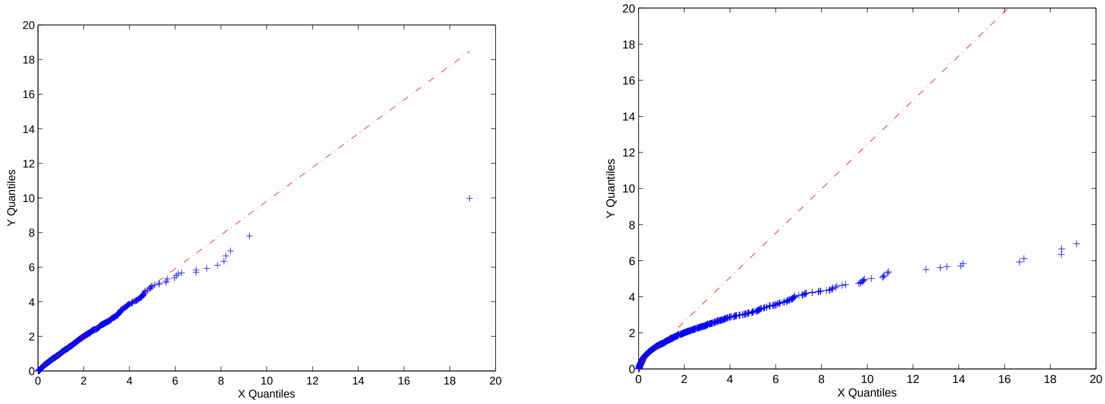

Figure 2: QQ-plots of integrated intensity (time-changed waiting times) against exponential, for Hawkes model (left) and Poisson model (right)

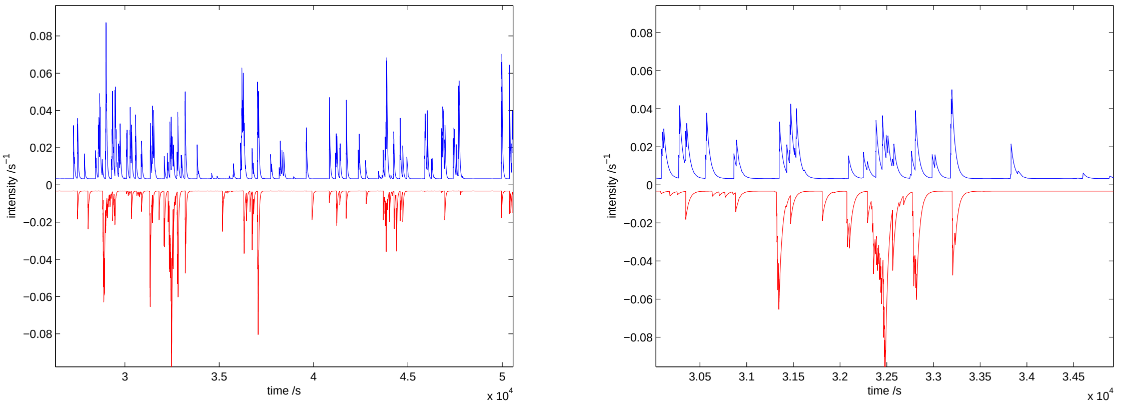

Figure 3: Behaviour of fitted intensities over a period of about a day (left) and about an hour (right). The intensity of selling is shown as negative for ease of interpretation.

4.3 Price impact function for fitted model

We now apply the results of Section 3 to the fitted model to calculate theoretical expected trade imbalance, covariance density and price impact function.

4.3.1 Prediction of trading intensity

We can use (19) to determine the expected total future imbalance of trades through time. Figure 4.3.1 shows the variation in expected imbalance over a half hour period. The highest expected imbalance in the sample was 12.4 trades and its time weighted mean absolute value was 0.55 trades. It should be stressed that *if* our price formation model (13) is correct, then there is no money to be made out of this predictability, as it is already impounded in the price.

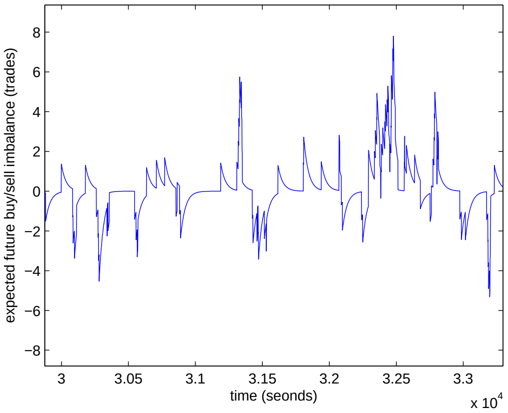

Figure 4: Expected total future imbalance of buy over sell trades, shown over a period of about half an hour

4.3.2 Covariance density

We calculate the conditional intensities of buys and sells following a buy using (11). We compare this fitted parametric estimate to the empirical histogram of conditional arrival rates in Figure 5. It seems that our model and/or estimation procedure has failed to pick up on a small cross-exciting effect. It also seems that the unconditional mean intensity of the fitted model is a bit too low. Neither of these problems affects the substance of what follows.

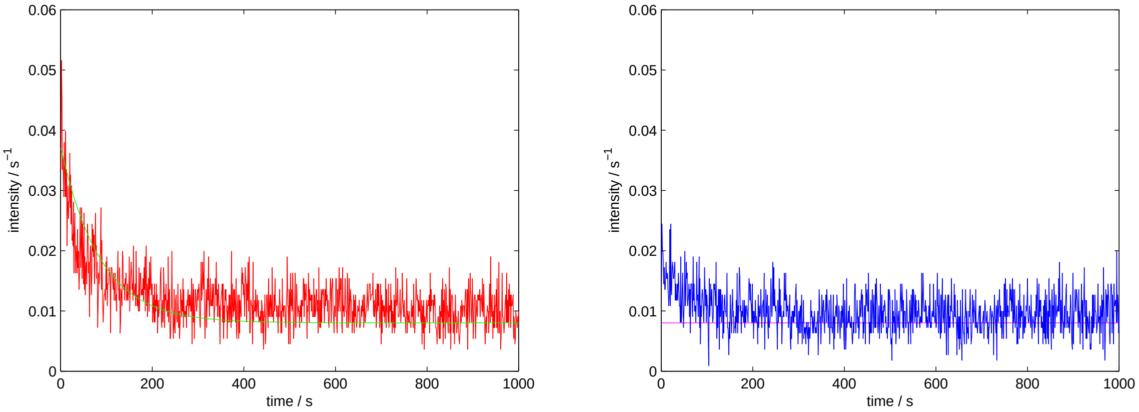

Figure 5: Conditional intensity of buys (left) and sells (right) following a buy, empirical (red/blue) and fitted (green/magenta)

4.3.3 Price impact function

Since $\alpha_{\text{cross}} = 0$, the price impact function has the form (22), with $\alpha = \alpha_{\text{same}}$, $\beta = \beta_{\text{same}}$, with θ to be determined. We show (Figure 6) the price impact function with θ normalised to one. The theoretical initial price impact of a trade is around 2.4 times its permanent impact. This ratio, equal to $\beta/(\beta - \alpha)$ in the model with no cross-excitement, is an important characteristic of the market. It can also be

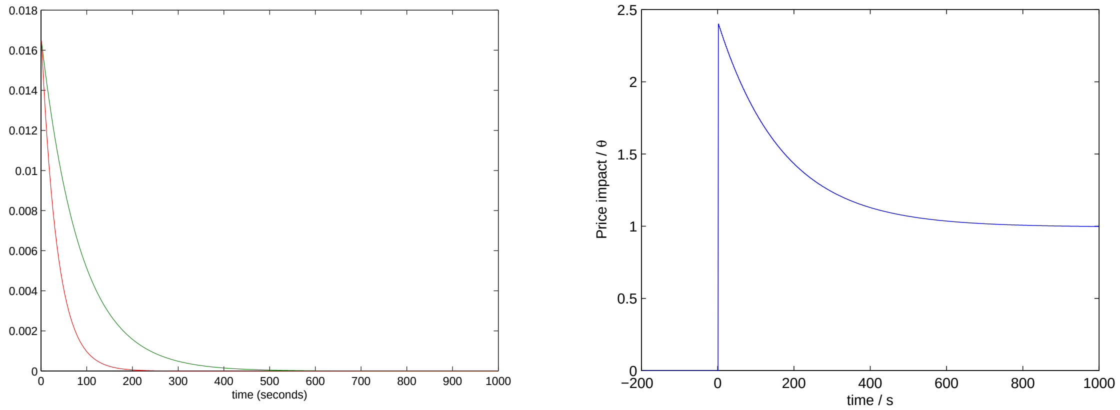

Figure 6: Direct (lower line) and total (upper line) impact on buy intensity of a buy (left); corresponding price impact function (right)

estimated in a trade time models via the autocorrelation function of buys and sells written as a sequence of 1s and −1s. A value of one would imply that a buy trade does not predict any further net imbalance of buys and sells; A value of less than one would imply that a buy trade tends to be followed by an imbalance of sells over buys. Preliminary investigations of other FX currency pairs indicate that 2.4 is relatively high but not atypically so.

5 Trading optimisation

We now turn to the problem faced by a trader in the market described above. We set out to model the trader's dilemma: trade too fast and be penalised by unnecessary market impact; trade too slowly and run the risk of adverse price movements before completing the trade. It is intuively clear that the shape of I determines the cost of trading fast relative to trading slowly: the greater the difference between $I(0)$ and $I(\infty)$, the greater the potential cost saving from waiting a long time between trades; and the more quickly I decays the less time we need to wait between trades so that most of the price impact is dissipated before we put in the next trade — we will formalise this below. The risk part of the model can also be extracted from the dynamics discussed above, but this is non-essential and we choose a more basic model of risk.

Throughout this section we assume that the trader measures performance relative to the mid price of the asset at the time he receives the instruction to trade. We assume further that he has an exponential utility function with risk aversion parameter γ, so that he aims to maximise $\mathbb{E}\overline{P} - \gamma\mathbb{V}\overline{P}$, where $\overline{P}$ is the average price realised for the trade[13].

[13] This requires the distribution of prices to be Normal; clearly price movements caused by trades are discrete and so the distribution is not Normal. More significantly, the volatility clustering inherent in self-exciting models implies that price movements would not even be Normal in a continuous time limit.

5.1 Modelling price as a controlled point process

We now discuss a trade path optimisation problem in the spirit of the one presented in [2]. A trader with initial inventory[14] Y_0 tries to maximise sale price penalised by some risk term if he holds inventory for too long. In our model, trading is discrete; we choose the units of Y_0 so that the the volume of each trade is one; we call this volume "standard trade size". Our trader is therefore required to sell his inventory in Y_0 discrete trades, at times $0 = \tau_1 \leq ... \leq \tau_{Y_0}$. For simplicity we consider the problem where $\tau_1 \leq ... \leq \tau_{Y_0}$ are to be determined at the outset of trading, so that we think of them as deterministic rather than random stopping times[15].

We assume bid price[16] dynamics of the form:

$$P_t = P_0 + \int_0^t I(t-u) dF_u + \int_0^t I(t-u)\, dY_u\,. \tag{25}$$

The second term represents others' trades, the third our own. Note that we employ the convention that if we need to sell, $Y_0 > 0$ so that $dY_u = -\sum_{n=1}^{Y_0} \delta(\tau_n)$, so the contribution of our trades to price is negative. Implicit in the application of the propagator I to both our trades and others' trades is the assumption that the market maker (i.e. the rest of the market) believes that the sum of our trades and everyone else's is a Hawkes process with the parameters estimated above.

Time $t = 0$ corresponds to the time our trader receives the instruction to trade. (25) does not include any history of others' trading before this time. We could incorporate a history of others' trading by evaluating the second term in (25) from some time in the past. If we did this, the *exact* optimal trading strategy would depend on the history of trades before time zero, but the approximation (28) below would eliminate this dependence.

5.1.1 Expected price impact

Consider the component of the price process that is not caused by our own trades:

$$Q_t = P_0 + \int_0^t I(t-u)\, dF_u\,. \tag{26}$$

By definition of I, Q is a martingale, although it has non-constant instantaneous variance due to the self-exciting nature of F. The expected price of a trade at time t then depends only on the effect of our own trading $\int_0^t I(t-u) dY_u$, so that

$$\mathbb{E}(Y_0 \overline{P}) = Y_0 P_0 - \sum_{i=1}^{Y_0} \sum_{j=1}^{i} I(\tau_i - \tau_j)\,. \tag{27}$$

5.1.2 Accounting for risk

In order to produce a convenient form for the risk term $\mathbb{V}\overline{P}$, we approximate Q by a random walk with *constant* volatility[17] σ. So, given our agent's *own* trading

[14] This position might have arisen from an investment bank's client, for example.

[15] Although we suspect that for the same reasons as outlined in [2], the solutions of the deterministic and stochastic optimisation problems coincide in the exponential utility case.

[16] Since the bid-mid spread is constant in this model, it does not really matter whether we model mid or bid. The spread can be thought of as an unavoidable cost.

[17] This removes the path-dependency of optimal selling strategy

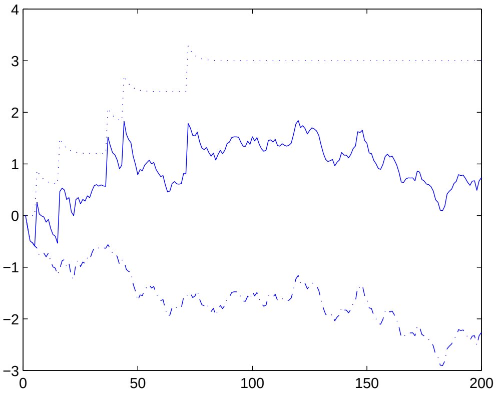

Figure 7: Linear model of controlled price process. The solid line is the price process, the dotted line the effect of our agent's trades and the mixed line a random walk representing the effects of others' trades. The units of time and price are arbitrary.

times $\tau_1, ..., \tau_{Y_0}$ we can write the price process as

$$P_t \approx P_0 + \sigma Z_t - \sum_{\tau_i < t} I(t - \tau_i)\,, \tag{28}$$

where Z_t is (for example) a standard Brownian motion[18], and σ is to be determined (or rather subsumed into the risk aversion coefficient). This model is illustrated in Figure 7. We then have

$$\mathbb{V}(P_t|P_0, \tau_1, ...\tau_{Y_0}) \approx \sigma^2 t\,, \tag{29}$$

so that for a trading strategy $\tau_1, ...\tau_{Y_0}$

$$\mathbb{V}(Y_0\overline{P}) \approx \mathbb{V}(\sum_{i=1}^{Y_0} P_{\tau_i}) = \sum_{i=1}^{Y_0}(Y_0 - i + 1)^2(\tau_i - \tau_{i-1})\,. \tag{30}$$

5.2 Optimisation

We then consider the usual trading mean-variance optimisation

$$\max_{\tau_1 < ... < \tau_{Y_0}} \mathbb{E}(\sum_{i=1}^{Y_0} P_{\tau_i}) - \gamma\mathbb{V}(\sum_{i=1}^{Y_0} P_{\tau_i})\,, \tag{31}$$

or equivalently[19]

$$\max_{\tau_1 < ... < \tau_{Y_0}} -\sum_{i=1}^{Y_0}\sum_{j=1}^{i} I(\tau_i - \tau_j) + \gamma\sigma^2\sum_{i=1}^{Y_0}(Y_0 - i + 1)^2(\tau_i - \tau_{i-1})\,. \tag{32}$$

Even for nice I such as the exponential for in (21), an analytic solution of (32) is not possible, but it is relatively easy to solve numerically, for example by a quasi-Newton method with a barrier function [17] to deal with the constraint $0 < \tau_1 < ... < \tau_{Y_0}$.

[18] The important thing is that Z should have independent increments with variance Δt.

[19] We are taking the optimal trading times τ_i to be deterministic.

5.2.1 Parameterisation for EUR/PLN

We now solve the test problem of selling ten standard lots in EUR/PLN. The average trade size in the market is of order €1m so ten lots corresponds to €10m. We scale price such that $\theta = 1$. Regressions indicate that in fact the permanent price impact of a trade in this market is 0.42 ticks[20], so our price unit θ can be interpreted as this amount. The one remaining parameter is $\gamma\sigma^2$, which we tune to show conservative, average and aggressive trading strategies. Practitioners can of course measure σ, and will have their own views about an appropriate value for γ.

5.3 Results

The results of solving (32) with these parameters are displayed below in two forms, for various values of $\gamma\sigma^2$, corresponding to a highly risk averse, typically risk averse (in a sense discussed below) and relatively risk-tolerant trader. The traders are assumed to start with an inventory of ten units at time zero. The top-left hand graph in Figure 8 shows their inventory as a function of time. The more risk-averse the trader, the more quickly they dispose of their inventory. The rest of the Figure shows the expected price impact of their trading strategies, and the average price per unit they expect to pay. We see that for the risk-averse, impatient trader, the price change is expected to overshoot its equilibrium value of minus ten, and the average price received is considerably less than this equilibrium value; for the "average" trader, the average price paid is around the equilibrium value; for the patient trader, the average price paid is less. In all cases, more than one unit of the trade is executed immediately at $t = 0$; this is discussed further below. For comparison, we also show the solutions of (32) for $Y_0 = 4$, with the same levels of risk aversion. For the two more risk-averse traders, these solutions coincide, and require immediate sale of all four units.

5.4 Discussion

Firstly note the recommendation of these models that a large chunk of the trade is executed immediately at the start of the trading period, followed by a stream of smaller units throughout the rest of the period. This is consistent with conversations with traders, and results reported in [27]. However, it also highlights and important shortcoming of our model: the linear specification for the price impact function (16), (22) carries with it the implication that price realised for a simultaneous trades of multiple units is a linear function of volume[21]. There is significant debate in the literature regarding the shape of this temporary price impact function (e.g. [21], [15]), which is related to the average shape of the liquidity profile in the limit order book. Without going into the details, (16) implies that the profile of standing liquidity is constant out to infinity. We do not currently have any data on the liquidity profile in this market, but common sense dictates that liquidity per price tick must become thinner after a certain distance from best bid or ask, so that linear instantaneous price impact must break down at some stage. We must therefore be very careful in applying the results of models such as this one, which may very

[20] The tick size is 0.0005 Zloty Euro^{-1}, and θ is measured to be 0.00021 Zloty Euro^{-1} Trade^{-1}.

[21] As can be seen by letting the inter-trade time tend to zero

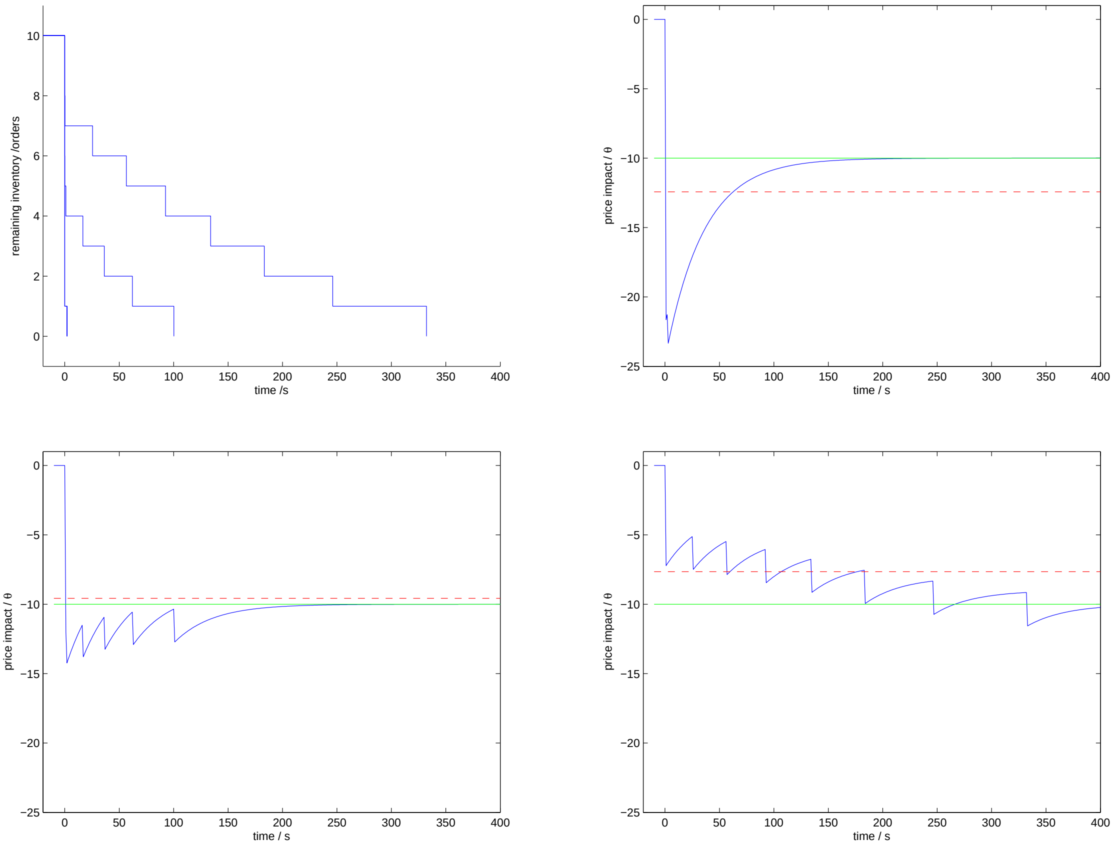

Figure 8: Top left: Schedules for selling ten units for three different levels of risk aversion: $\gamma\sigma^2 = .1$ (lower line),$\gamma\sigma^2 = .01$ (middle line), $\gamma\sigma^2 = .001$ (top line); expected price impact of these three strategies over time, showing average price paid (dotted flat line) and eventual price impact (solid flat line) - $\gamma\sigma^2 = .1$ (top right), $\gamma\sigma^2 = .01$ (bottom left), $\gamma\sigma^2 = .001$ (bottom right)

easily recommend an initial trade of a size which is either impossible or can only be completed at a *very* unfavourable price.

There are a number of reasons not explored in our simple model why the market might not offer linear instantaneous price impact. These include richer information effects, market-maker risk aversion and inefficient or uncompetitive market-making. An alternative suggestion is that the relationship between volume and expected future trades is itself nonlinear. Whether a non-linear version of (16), derived for example from a nonlinear self-exciting process, could be squared with the instantaneous price impact function implied by observed limit order books is an interesting area for future research.

We now turn to the question of how well our traders have done in disguising their trades. If they were obliged to disclose their position to the market maker before commencing trading, they would realise a (normalised) price impact cost of N. We first consider $N = 10$; noting that the average cluster size in this market is 2.4 trades, trades of this size are probably relatively rare. Intuitively we would expect to be able to complete the first few tranches of such a trade cheaper than the final equilibrium price, because the market does not expect a cluster of this

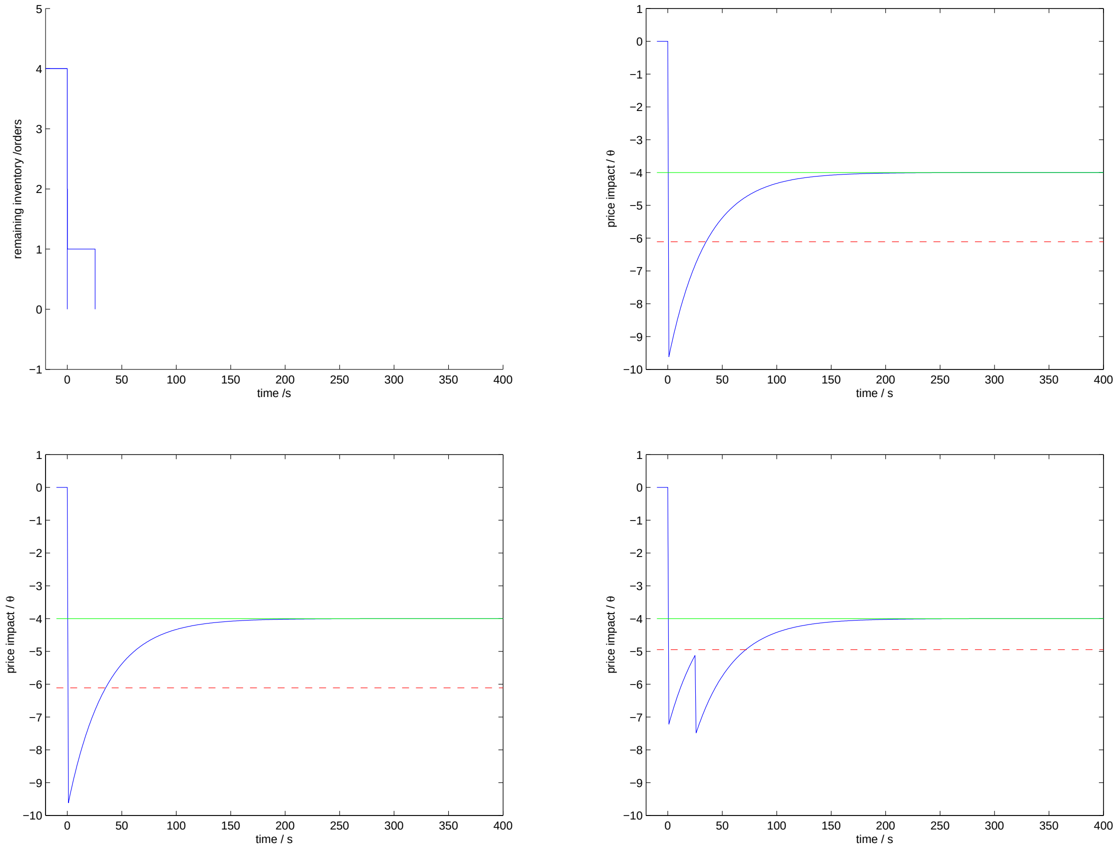

Figure 9: Top left: Schedules for selling four units for three different levels of risk aversion: $\gamma\sigma^2 = .1$ (left hand line),$\gamma\sigma^2 = .01$ (left hand line also), $\gamma\sigma^2 = .001$ (right hand line); expected price impact of these three strategies over time, showing average price paid (dotted flat line) and eventual price impact (solid flat line) - $\gamma\sigma^2 = .1$ (top right), $\gamma\sigma^2 = .01$ (bottom left), $\gamma\sigma^2 = .001$ (bottom right)

size. This is the case in our model — the first four trades do not push the price above the final price even if done simultaneously. Even the most risk-averse trader, who completes nine out of ten trades immediately and the last trade a second later, does not pay hugely over the odds — the market-maker takes a small profit; the least risk-averse trader manages to transact at a discount to fair price, causing the market-makers to lose money. For $N = 4$, i.e. slightly larger than the typical cluster size, we see that even for the lowest level of risk aversion investigated, the trader pays over the odds, and the market-maker realises a profit. In fact, we must set $\gamma = 0.0001$ for the trader realise the final price on average, and the market-maker break even (not shown in graphs). For $N \leq 3$, however patient he is, a trader will always pay over the odds[22].

We now consider whether a population of traders behaving as prescribed by our model could (at least approximately) generate the kind of point process we have assumed. Let us suppose for the sake of exposition that the sole reason

[22] The limiting average price as time between trades tends to infinity in the case $N = 3$ is $(2.4 + 3.4 + 4.4)/3 = 3.4 > 3$

for the observed clustering of trades is the splitting of large trades into smaller tranches[23]. Then by solving the traders' problem in the manner described above, and aggregating the actions of a population of heterogeneous traders, one could arrive at aggregate order arrival dynamics. Key parameters in such a model are the distribution of size of trades to be split and trader risk-aversion, which governs speed of execution. If these dynamics were exactly as assumed in the original solution of the market-makers' problem, we would have a game-theoretic solution to the interaction between market-makers and order-splitting traders.

We stress again that our paper does not purport to offer such a game-theoretic solution of the interaction between traders and market-makers. Putting aside the obvious difficulties in solving such a game, in order to parameterise it we would need to observe or infer the statistical properties of the client orders arriving at dealing desks in the market, as well as distribution of risk aversion parameters for the traders executing these orders. Even if we could do this, it would not be helpful if some of the traders were acting sub-optimally. Our methodology has been to subsume all of these concerns into the tractable yet flexible class of linear self-exciting processes.

Because of the limitations of linearity and the exponential parametric form, our market-maker is not as clever as he might be — he cannot detect nonlinear patterns and even some linear patterns such as periodicities[24] that may be present in the trade arrivals data. We do not know at the moment how financially significant such patterns might be to market-makers.

6 Summary and Future Work

In this paper we outlined how a rational risk-neutral market-maker would react to a linear self-exciting market order arrivals process. We showed that the price impact function is also linear, and that closed forms are available when the self-excitement function is exponential. These theoretical price impact functions were investigated in the context of a model for executing a large trade under risk aversion.

In the future we plan to investigate whether introducing nonlinearities in the trade arrival intensities results in a better fit to the data. If so, we will investigate the theoretical form of price impact of a trade for such nonlinear models, and apply this to the trade splitting optimisation model. We also plan to investigate non-exponential forms for the self-excitement function in the linear case, including the possibility of long memory effects, and their consequences for price impact function and trade splitting optimisation. We hope to obtain data on volumes shortly, and investigate the effect of trading volume "marks" in the Hawkes model of arrivals.

Another line of possible work concerns whether, and by what mechanism, real limit order book markets implement the price impact behaviour outlined. If they do not, it may indicate an inefficiency that can be exploited in detailed microstructural trading algorithms. If market-makers are able, in aggregate, to implement

[23] [16] present a model linking order splitting strategies to the autocorrelation of the sequence of buy and sell trades (which works is trade time, but the ideas would apply in calendar time with minor adjustment)

[24] An simple example of periodicity from the early days of program trading in equities: early order-splitting engines sometimes traded equal amounts at regularly spaced intervals during the day. Market players then developed "sniffers" based on spectral analysis to detect and take advantage of these patterns.

effectively the linear filter described in this paper, we would like to establish the institutional and / or behavioural features that allow them to do so.

References

[1] R.A. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10:1–18, 2003.

[2] R.A. Almgren and N. Chriss. Optimal execution of portfolio transactions. *J. Risk*, 3:5–39, 2000.

[3] J-L. Bauwens and N. Hautsch. Dynamic latent factor models for intensity processes. *UCLA CORE discussion paper*, 2003.

[4] D. Bertsimas and A. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1:1–50, 1998.

[5] Jean-Philippe Bouchaud, Yuval Gefen, Marc Potters, and Matthieu Wyart. Fluctuations and response in financial markets: the subtle nature of 'random' price changes. Technical Report 0307332, Science & Finance, Capital Fund Management, July 2003. available at http://ideas.repec.org/p/sfi/sfiwpa/0307332.html.

[6] C. Bowsher. Modelling security events in continuous time: Intensity based, multivariate point process models. *Nuffield College Economics Discussion Papers*, 2005W-26, 2005.

[7] C. Bowsher. Modelling security events in continuous time: Intensity based, multivariate point process models. *Nuffield College Economics Discussion Papers*, 2002W-22, 2005.

[8] Daley D. and Vere-Jones D. *An Introduction to the Theory of Point Processes*. Springer-Verlag, 2003.

[9] R.A. Davis, T.H. Rydberg, N. Shephard, and Streett S.B. The cbin model for counts. *Unpublished discussion paper*, 2001.

[10] A. Dufour and R.F. Engle. Time and the price impact of a trade. *J. Finance*, 55:2467–2498, 2000.

[11] D. Easley and M. O'Hara. Time and the process of security price adjustment. *J. Finance*, 47:577–605, 1992.

[12] R.F. Engle and Lunde A. Trades and quotes: a bivariate point process. *J. Financial Econometrics*, 1:159–198, 2003.

[13] R.F. Engle and J.R. Russell. Forecasting frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model. *J. Empirical Finance*, 4:187–212, 1997.

[14] R.F. Engle and J.R. Russell. Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 66:1127–1162, 1998.

[15] J.D. Farmer, F. Lillo, and R.N. Mantegna. Econophysics: Master curve for price impact function. *Nature*, 421:129–30, 2003.

[16] J.D. Farmer, F. Lillo, and S. Mike. Theory for long memory in supply and demand. *Phys Rev E*, 6(2):287–297, 2005.

[17] R. Fletcher. *Practical Methods of Optimization.* Wiley, 1987.

[18] K. Giesecke and L. Goldberg. A top down approach tomulti-name credit. *Available at SSRN: http://ssrn.com/abstract=678966*, 2005.

[19] L. Glosten. Components of the bid-ask spread and the statistical properties of transaction prices. *Journal of Fiance*, 42:1293–1307, 1987.

[20] L. Glosten and P. Milgrom. Bid, ask, and transaction prices in a specialist market with heterogeneously informed agents. *Journal of Financial Economics*, 14:71–100, 1985.

[21] Joel Hasbrouck. Measuring the information content of stock trades. *Journal of Finance*, 46(1):179–207, 1991.

[22] Joel Hasbrouck. Trading fast and slow: Security market events in real time. New York University, Leonard N. Stern School Finance Department Working Paper Seires 99-012, New York University, Leonard N. Stern School of Business-, February 1999. available at http://ideas.repec.org/p/fth/nystfi/99-012.html.

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

[24] A.G. Hawkes and D. Oakes. A cluster representation of a self-exciting point process. *J. Applied Probability*, 11:493–503, 1974.

[25] Large J. Measuring the resiliency of an electronic order book. *Journal of Finance*, forthcoming, 2005.

[26] A. MacNeil. Self-exciting processes for extremes in financial time series. *ETHZ Working Paper*, 2005.

[27] A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. *MIT working paper, 2004*, 2004.

[28] Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30, 2A:243–261, 1978.