Duration and Order Type Clusters
When Traders React and on which Market Side

Wing Lon NG
CAWM - WWU Muenster
Institut für Ökonometrie und Wirtschaftsstatistik
Am Stadtgraben 9, 48143 Münster, Germany
05wing@wiwi.uni-muenster.de

27th February 2004

Abstract

This paper introduces an extended bivariate autoregressive conditional duration (ACD) framework for modelling the arrival process of buy and sell orders in a limit order book. The model contains two dynamic components to capture the observed clustering of durations and limit order types: a duration process combined with a new logistic "order-type" process, both depending on a common natural filtration. It can be manifested that the state of the order book as well as the success and the speed of the matching process have a significant influence on the bid/ask quotes, and thus, affect the traders´ decisions when and on which side of the market to trade.

Key Words: Ultra high frequency transaction data, limit order book, market microstructure, ACD model, dynamic logit model, bivariate point process.

JEL Classifications: C14, C22, C32, C41.

1 Introduction

There is a large theoretical and empirical literature on the microstructure of financial markets, boosted by the increased availability of ultra high frequency transaction data. These time stamped data are characterized by one main feature: the irregularity of time intervals between two observations. As the time variable is considered as stochastic, the study of financial econometric models requires an alternative method to ordinary (fixed-) time series analysis. Based on the influential work by Engle and Russell (1998) and Engle (2000) who successfully modelled this specific time structure, many

studies have concentrated on the further improvement of autoregressive conditional models, especially duration (ACD) and intensity (ACI) processes, in order to describe the order book activities more accurately. Since the duration between transactions and the timing of a transaction itself heavily affect the traders´ ordering decisions, they become important variables explaining the development of intraday returns in financial markets. As illustrated in Engle (2000) and Engle and Lunde (2003), the stochastic properties of the trade arrival process and, in particular, their durations are a decisive reason for volatility. But only examining the trade and its impact on prices and returns is not enough: recent studies demonstrated the importance of the **quote**´s timing and information content. It is often neglected that financial electronical markets are designed for a rapid matching of buyers and sellers of assets. Therefore, the statistical analysis of the dynamic market process must incorporate the distribution of waiting times and the arrival frequency of incoming orders. Whereas transaction data only mirror the state of the order book at the intersection of the supply and demand side, quotes allow us a deeper insight into the market participants´ prior intentions to trade.

A few recent studies explore electronic order books, taking a closer look at the timing and the content of the quotes. Hall, Hautsch and MacCulloh (2003), for example, run a probit regression to extract the factors driving the traders´ bid and ask decisions. Modelling the joint intensity of the buy and sell arrival process, they show in their empirical results that the state of the order book has a significant influence on the bid and ask intensity. Although their intensity approach is convenient for multivariate specifications and time varying covariates, it is far less intuitive and forecasts are computationally burdensome (see also Russell (1999), Bowsher (2002), Bauwens and Hautsch (2003)). In contrast, Engle and Lunde (2003) use two time scales in their analysis. They treat the arrival of trades and succeeding quotes as a bivariate, dependent point process. The arrival of each event type is influenced by the past history of both processes and other market information. Due to the combination of trade and quote data, a complicated situation arises. This makes the specification of the dependence between duration pairs very difficult and clearly shows that the common ACD model has its limits: the weakness of this kind of models is that they are not suitable for multivariate specifications, because one must condition on the information available at the beginning of each duration. In contrast to an intensity process, it is difficult (in a multivariate duration process) to take into account new information during the actually lasting waiting time.

This paper solves this problem by introducing an extended ACD model considering **all** points of both the trade and quote processes without any distinction between bid-, ask- and trade durations, as shown in figure 1. Since transactions are always initiated by either an ask or a bid limit order (in the continuous trading phase), it is sufficient only to record the arrival times of all incoming orders and their type. A simple Generalized Gamma-

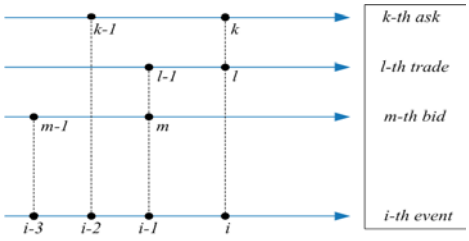

Figure 1: All event points

ACD(2,1) model was used to capture the empirically often observed duration clusters. Additionally, to restore the respective type of a limit order and its contents, an innovative "dynamic logistic process" is affixed to the ACD model. A further advantage of this model is that it does not only allow the prediction of the next order type given the past history, but also easily solves the "zero-duration" problem that often occurs, when, for example, a high demand cannot be satisfied by one single supplier and, therefore, must be divided into n transactions, all executed at one point of time, as shown in figure 2. In univariate models, the "zero-duration" problem is often eluded by aggregating n transactions or, worse, eliminating $n-1$ transactions.

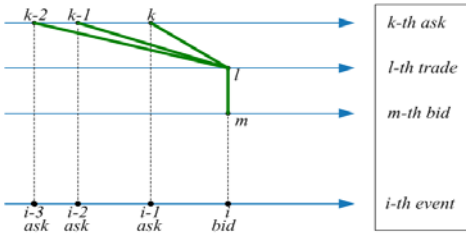

Figure 2: Aggregated transactions

Contrary to Engle and Lunde (2003), where two duration process are modeled jointly, the model suggested in this paper is much easier to handle: there is (a) one duration process that captures the whole time structure and (b) an "order-type" process containing the remaining information of the order book. Technically speaking, it is like Pohlmeier and Liesenfeld´s Integer Count Hurdle model (2003) or Engle and Russell´s Autoregressive Conditional Multinomial model (2002), both decomposing the general transaction price process into binary processes indicating the size and the direction of price changes and a count process for the size of the price change conditioned on the direction of the price change.

The main objective of this paper is to investigate the arrival process of bid and ask limits and to discover what determines the traders' decisions

when and on which side of the market to trade. To achieve this aim, the author jointly models the dynamics of the duration process of all time-stamped event arrivals in the order book as well as its influences on the stochastics of the order type process. As Engle and Lunde found out, quotes and trades tend to cluster in time in both a deterministic and stochastic way. In comparison to other approaches, this bivariate model is easier to compute and to estimate due to its linear AR structure interlocked twice in the specification.

The outline of this article is as follows: In section 2 the model is introduced and described. Section 3 shows the smoothing technique and discusses the ML estimation procedure. In section 4 the data and the empirical results will be presented, especially with respect to the economic implications. Section 5 concludes.

2 The Extended ACD-Model

As high frequency data arrive in irregular time intervals, researchers are concerned not only with the variable of genuine interest (i.e. price, quote, volume), but also with the arrival time of each event. Generally, transaction data can be described by two types of random variables. The first one is the time T of the transaction and the other one is the observation $\mathbf{Z}$ (called marks) linked with T. Consider the arrival times

$$t_0, t_1, t_2, ...$$

with $t_i \in \mathbb{R}^{\geq 0} \forall i$, as random variables distributed in time by a point process. Here $(t_i)_{i \in \mathbb{N}}$ is the sequence of arrival times of an incoming order, not necessarily a transaction. (When and why an order initiates the execution of one or more transactions will be discussed later.) It is convenient to introduce a counting function $N(t)$ which simply indicates the number of event arrivals that have occurred at or prior to time t. This will be a monoton-increasing step function with unit increments at each arrival time. Obviously, $N(t)$ is a simple jump process with $N(t_0) = 0$. Further, define the filtration

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}_{i-1} \subseteq \mathcal{F}_i = \sigma(t_0, t_1, ..., t_i)$$

with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Thus $\mathcal{F}_i$ is the σ-field generated by all time random variables observed till t_i. The instantaneous probability of an event at t is called the intensity of the process. In time dependent processes this intensity is obtained by conditioning on past information. Define the conditional intensity of a process as

$$\lambda\left(t|N(t), t_1, t_2, ..., t_{N(t)}\right) = \lim_{\Delta t \to 0} \frac{P\left(N(t+\Delta t)|N(t), t_1, t_2, ..., t_{N(t)}\right)}{\Delta t}.$$

This function provides a complete description of the point process´ full dynamics and is similar to the hazard rate, which is often used in survival analysis and technometrics (see, for example, Cox and Oakes (1984) or Lancaster (1990)). Now let

$$X_i = t_i - t_{i-1}$$

with $t_0 = 0$, where X_i is the i-th duration between the i-th and $(i-1)$-th incoming order. The crucial assumption for ACD models is that the dependence structure can be summarized by one function, namely the conditional expected duration Ψ_i, adapted to the filtration $\mathcal{F}_{i-1}$. Therefore let

$$\begin{aligned} E\left(X_i|\mathcal{F}_{i-1}\right) &\equiv \Psi\left(X_i|X_{i-1},...,X_1;\boldsymbol{\theta}_1\right) \\ &= \Psi_i \\ &= \omega + \alpha_1 X_{i-1} + ... + \alpha_p X_{i-p} + \beta_1 \Psi_{i-1} + ... + \beta_q \Psi_{i-q} \\ &= \omega + \sum_{j=1}^{p} \alpha_j X_{i-j} + \sum_{k=1}^{q} \beta_k \Psi_{i-k}. \end{aligned}$$

where the parameters $\omega, \alpha_1, ..., \alpha_p, \beta_1, ..., \beta_q$ are all included in $\boldsymbol{\theta}_1$. It is clear that the probabilistic structure of the conditional duration Ψ_i is similar to that of a GARCH process and, hence, this class of models are also called "*autoregressive conditional*", characterized by the lag length of the past durations *(ACD(p,q))*. Such as in an Accelerated Failure Time model it is now assumed that

$$X_i = \Psi_i \cdot \varepsilon_i$$

with

$$\varepsilon_i \overset{i.i.d.}{\sim} f\left(\varepsilon_i|\mathcal{F}_{i-1};\boldsymbol{\theta}_1\right) = f\left(\varepsilon_i;\boldsymbol{\theta}_1\right).$$

The main property here is that the errors $\varepsilon_i = \frac{X_i}{\Psi_i}$ are independent and identically distributed random variables with the probability density function $f(.)$, which must be specified. While Engle (2000) preferred an Exponential or a Weibull distribution, other authors favoured more flexible alternatives like the Burr- or F-distribution (Fernandes and Grammig (2000), Hautsch (2002)). Of course, this density must have a non-negative support. In this paper a generalized gamma distribution

$$\begin{aligned} f_{GGamma}\left(\varepsilon_i;\boldsymbol{\theta}_1\right) &= f_{GGamma}\left(\frac{X_i}{\Psi_i};\boldsymbol{\theta}_1\right) \\ &= \frac{\gamma}{X_i \cdot \Gamma(\lambda)} \left(\frac{X_i}{\Psi_i} \cdot \frac{\Gamma\left(\lambda + \frac{1}{\gamma}\right)}{\Gamma(\lambda)}\right)^{\gamma\lambda} \\ &\quad \cdot \exp\left(-\left(\frac{X_i}{\Psi_i} \cdot \frac{\Gamma\left(\lambda + \frac{1}{\gamma}\right)}{\Gamma(\lambda)}\right)^{\gamma}\right) \end{aligned}$$

is used, where λ is the shape parameter and γ the scope parameter of the density function (both included in $\boldsymbol{\theta}_1$). Further assume that ε_i are independent of X_i. Since the durations and expected durations are positive, the multiplicative disturbance naturally will have positive probability only for positive values and it must have a mean of unity

$$\begin{aligned} E(\varepsilon_i) &= 1 \\ Var(\varepsilon_i) &= \sigma_\varepsilon^2. \end{aligned}$$

This assumption requires all temporal dependence of the durations to be captured entirely by the mean function. This hypothesis is testable in practice by using the standardized durations. Usually different duration models for transaction data are developed via the dependence of the conditional expectation on the past durations. Hence, several new types of ACD models can be created by varying the the functional form $g(.)$ of the conditional mean equation:

$$\begin{aligned} X_i &= g(\Psi_i) \cdot \varepsilon_i \\ &= g(E(X_i|X_{i-1}, ..., X_1; \boldsymbol{\theta}_1)) \cdot \varepsilon_i \end{aligned}$$

Bauwens and Giot (2001), for example, introduced the Log-ACD with two possible modifications. Certainly, alternative nonlinear dependence structures are also possible (Fernandes and Grammig (2000), Bauwens (2000)).

The simple ACD model as the most common approach accommodates duration clustering through the time dependency of durations and represents in its simplest form a time series model of time, making it relatively easy to understand. It is a dynamic point process model in which the conditional expectation is written as a linear function of past durations. But often there are additional observations $\mathbf{Z}_i = (Z_{1,i}, ..., Z_{m,i})$ associated with the arrival times t. For financial transaction data, a plethora of information is linked with the time stamps (including price, volume, bid and ask quotes, depth, etc.). In this case, the new point process $(t_i, \mathbf{Z}_i)_{i \in \mathbb{N}}$ will become "marked". Depending on the economic question at hand, either the arrival time, or the marks, or both may be of interest. Since the marks associated with the i-th arrival time are not included in $\mathcal{F}_i$, define a more comprehensive filtration $\mathcal{F}_i^*$ representing a σ-field that contains all past arrival times and marks till t_i

$$\mathcal{F}_i \subset \mathcal{F}_i^* = \sigma(t_0, t_1, ..., t_i; \mathbf{Z}_0, \mathbf{Z}_1, ..., \mathbf{Z}_i) \qquad \forall i.$$

The ACD model is modified by including the marks $\mathbf{Z}_i$ in the mean equation in order to model the time structure more accurately. However, it turns out that the linear specification

$$\Psi_i = \omega + \sum_{j=1}^{p} \alpha_j X_{i-j} + \sum_{k=1}^{q} \beta_k \Psi_{i-k} + \sum_{l=1}^{r} \sum_{w=1}^{m} \tau_w Z_{w,i-l}$$

is insufficient, as the order types will be included in the vector of marks. To analyze the cluster structure of order types, one needs a (non-linear) function $\Lambda(.)$ that is able to describe the type of the incoming limit orders at first. Therefore, denote

$$\begin{aligned} Y_i &= i\text{-th order type signaling the market side of the trader} \\ &= \begin{cases} 0 & \text{if order = ask-limit} \\ 1 & \text{if order = bid-limit} \end{cases}. \end{aligned}$$

Obviously $(t_i, Y_i)_{i\in\mathbb{N}}$ is also a marked point process. Since Y_i is a dummy variable representing the market side at t_i, one has to model its binary marks by their respective probabilities. (Besides, it is numerically difficult to differentiate between one "zero" observation and a longer zero-string, which means that the cluster structure would get lost if one uses a count model.) Assume that the probability of a bid-order conditional on $\mathbf{Z}_{i-1}$is given by the following logit model

$$\begin{aligned} P(Y_i = 1|\mathbf{Z}_{i-1}) &\equiv \Lambda(\mathbf{Z}_{i-1}\boldsymbol{\tau}) \\ &= \frac{\exp(\mathbf{Z}_{i-1}\boldsymbol{\tau})}{1+\exp(\mathbf{Z}_{i-1}\boldsymbol{\tau})} \end{aligned}$$

with

$$\begin{aligned} \mathbf{Z}_{i-1}\boldsymbol{\tau} &= \sum_{w=1}^{m} \tau_w Z_{w,i-1} \\ &= \tau_1 Z_{1,i-1} + \tau_2 Z_{2,i-1} + ... + \tau_m Z_{m,i-1}. \end{aligned}$$

Here, $\Lambda(.)$ is the distribution function of the standard logistic distribution, often used in panel data and microeconometrics. Moreover, from the probability of an ask order

$$\begin{aligned} P(Y_i = 0|\mathbf{Z}_{i-1}) &= 1 - \Lambda(\mathbf{Z}_{i-1}\boldsymbol{\tau}) \\ &= \frac{1}{1+\exp(\mathbf{Z}_{i-1}\boldsymbol{\tau})} \end{aligned}$$

one may derive (see Johnson, Kotz and Balakrishnan (1995))

$$\begin{aligned} P(Y_i = y_i|\mathbf{Z}_{i-1}) &= P(Y_i = 1|\mathbf{Z}_{i-1}) \cdot P(Y_i = 0|\mathbf{Z}_{i-1}) \\ &= \Lambda(\mathbf{Z}_{i-1}\boldsymbol{\tau}) \cdot [1 - \Lambda(\mathbf{Z}_{i-1}\boldsymbol{\tau})] \\ &= f_{Logistic}(\mathbf{Z}_{i-1}\boldsymbol{\tau}) \end{aligned}$$

which is very important for the model´s joint density later on. The statistical problem is to estimate the probability of an order type dynamically, which requires (a) to specify the stochastic process of their arrival times, (b) to estimate all parameters recursively and then (c) to compute the likelihood function. To reconstruct the original structure of the order book

more accurately, as shown in figure 2, it is recommended to generate certain indicators measuring the temporal distance and the state of the order queue. Therefore, first introduce a new integer variable C_i summarizing the number of asks (bids) at time t_i since the last bid (ask). It is obvious that C_i is a right-continuous counting process $N\left(t_i^{Type}\right)$, cumulating the number of clustering orders of the same type on each market side until t_i

$$\begin{aligned} C_i^{Bid} &= N\left(t_i^{Bid}\right) \\ &= \begin{cases} 0 & \text{if last order is ask-initiated} \\ C_{i-1}^{Bid}+1 & \text{if last order is bid-initiated} \end{cases} \end{aligned}$$

$$\begin{aligned} C_i^{Ask} &= N\left(t_i^{Ask}\right) \\ &= \begin{cases} 0 & \text{if last order is bid-initiated} \\ C_{i-1}^{Ask}+1 & \text{if last order is ask-initiated} \end{cases} \end{aligned}$$

In case of an order type alteration, this counting variable will be reset to zero for the corresponding side (although there could be more unmatched orders since the last transaction). As a proxy for the buyers´ and sellers´ trading intensity, its aim is to measure the length of the actually queueing bid and ask limits in the order book. To display the temporal distance between the order types, one can introduce two new interesting duration variables

$$\begin{aligned} Dur_i^{Ask} &= \text{(cumulated) waiting time since the last ask order} \\ &= \begin{cases} X_i & \text{if last order is also ask-initiated} \\ Dur_{i-1}^{Ask}+X_i & \text{if last order is bid-initiated} \end{cases} \end{aligned}$$

$$\begin{aligned} Dur_i^{Bid} &= \text{(cumulated) waiting time since the last bid order} \\ &= \begin{cases} X_i & \text{if last order is also bid-initiated} \\ Dur_{i-1}^{Bid}+X_i & \text{if last order is ask-initiated} \end{cases} . \end{aligned}$$

Certainly, one could include additional regressors into the logit model (the real length or the cumulated volume of the bid/ask order queue, the volume of trades or the bid–ask spread, etc).

To model the order-type clusters, the order type probability is conditioned on the natural filtration $\mathcal{F}_i^*$ in a recursive manner, similar to the idea of GARCH and ACD. To emphasize the analogy of this "*autoregressive conditional logit*" model with the common ACD(p,q) , denote a general

ACL(u,v) model as

$$\begin{aligned} P\left(Y_i = 1|\mathcal{F}^*_{i-1}\right) &\equiv \Lambda\left(Y_i|Y_{i-1},...,Y_1,\mathbf{Z}_{i-1},...,\mathbf{Z}_1;\theta_2\right) \\ &= \Lambda_i \\ &= \Lambda\left(\sum_{j=1}^{u}\alpha_j' Y_{i-j} + \sum_{k=1}^{v}\beta_k'\Lambda_{i-k} + \sum_{l=1}^{r}\sum_{w=1}^{m}\tau_w Z_{w,i-l}\right) \\ &= \left[1+\exp\left(-\left(\sum_{j=1}^{u}\alpha_j' Y_{i-j} + \sum_{k=1}^{v}\beta_k'\Lambda_{i-k} + \sum_{l=1}^{r}\sum_{w=1}^{m}\tau_w Z_{w,i-l}\right)\right)\right]^{-1}. \end{aligned}$$

Of course, one may add further covariates to improve the model´s fit or use other suitable distribution functions. In this paper the following simple ACL(1,1) specification is considered

$$\begin{aligned} P\left(Y_i = 1|\mathcal{F}^*_{i-1}\right) &= P\left(Y_i = 1|\mathcal{F}^*_{i-1};\boldsymbol{\theta}_2\right) \\ &\equiv \Lambda_i \\ &= \Lambda\left(\alpha_1' Y_{i-1} + \beta_1'\Lambda_{i-1} + \mathbf{Z}^*_{i-1}\boldsymbol{\tau}\right) \end{aligned}$$

with

$$\mathbf{Z}^*_{i-1}\boldsymbol{\tau} = \tau_1 C^{Ask}_{i-1} + \tau_2 C^{Bid}_{i-1} + \tau_3 Dur^{Ask}_{i-1} + \tau_4 Dur^{Bid}_{i-1}$$

and the parameter vector $\boldsymbol{\theta}_2 = \left(\alpha_1', \beta_1', \tau_1, \tau_2, \tau_3, \tau_4\right)$. It should be stressed that Λ_i is a time varying probability of the order type.

The ACL model and the ACD model are intertwined as follows

$$\begin{aligned} X_i &= \Psi^*_i \cdot \varepsilon_i \\ &= \left(\Psi_i + \delta_1\Lambda^{Bid}_{i-1} + \delta_2\Lambda^{Ask}_{i-1}\right)\cdot\varepsilon_i \\ &= \left(\omega + \sum_{j=1}^{p}\alpha_j X_{i-j} + \sum_{k=1}^{q}\beta_k\Psi_{i-k} + \delta_1\Lambda^{Bid}_{i-1} + \delta_2\Lambda^{Ask}_{i-1}\right)\cdot\varepsilon_i \end{aligned}$$

and therefore

$$\begin{aligned} E\left(X_i|\mathcal{F}^*_{i-1}\right) &\equiv \Psi\left(X_i|X_{i-1},...,X_1;\Lambda_{i-1},...,\Lambda_1;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right) \\ &= \Psi^*_i \\ &= \omega + \sum_{j=1}^{p}\alpha_j X_{i-j} + \sum_{k=1}^{q}\beta_k\Psi_{i-k} + \delta_1\Lambda^{Bid}_{i-1} + \delta_2\Lambda^{Ask}_{i-1} \\ &= \omega + \sum_{j=1}^{p}\alpha_j X_{i-j} + \sum_{k=1}^{q}\beta_k\Psi_{i-k} + \delta_1\Lambda_{i-1} + \delta_2\left(1-\Lambda_{i-1}\right) \\ &= \underbrace{\left(\omega+\delta_2\right)}_{\omega'} + \sum_{j=1}^{p}\alpha_j X_{i-j} + \sum_{k=1}^{q}\beta_k\Psi_{i-k} + \underbrace{\left(\delta_1-\delta_2\right)}_{\delta}\Lambda_{i-1} \end{aligned}$$

with

$$\Lambda_i = \left[1 + \exp\left(-\left(\sum_{j=1}^{u} \alpha_j' Y_{i-j} + \sum_{k=1}^{v} \beta_k' \Lambda_{i-k} + \sum_{l=1}^{r}\sum_{w=1}^{m} \tau_w Z_{w,i-l}\right)\right)\right]^{-1}.$$

And of course, one can consider other nonlinear functional forms of the mean equation Ψ_i^*. To keep the computational burden acceptable, this paper concentrates on a GGamma-ACD(2,1) combined with an ACL(1,1). This bivariate model in its simplest form already contains a parameter vector $\boldsymbol{\theta}$ of 13 elements

$$\Psi_i^* = \omega' + \alpha_1 X_{i-1} + \alpha_2 X_{i-2} + \beta \Psi_{i-1}^* + \delta \Lambda_{i-1} \tag{1}$$

and

$$\begin{aligned} \Lambda_i &= \Lambda(\alpha_1' Y_{i-1} + \beta_1' \Lambda_{i-1} + \tau_1 C_{i-1}^{Ask} \\ &\quad + \tau_2 C_{i-1}^{Bid} + \tau_3 Dur_{i-1}^{Ask} + \tau_4 Dur_{i-1}^{Bid}). \end{aligned} \tag{2}$$

In this new bivariate model, the main dependence structure is captured by (1) and (2), each containing and influencing the information for the other process, both adapted to the filtration $\mathcal{F}_i^*$. The joint distribution is specified as

$$F_{X_i,Y_i}(x_i, y_i) = F_{X_i|Y_i=y_i,\mathcal{F}_{i-1}^*}(x) \cdot F_{Y_i=y_i|\mathcal{F}_{i-1}^*}(y)$$

that leads to the following interesting mixed density function

$$\begin{aligned} f_{X_i,Y_i}(x_i, y_i) &= f_{X_i|Y_i=y_i,\mathcal{F}_{i-1}^*}(x) \cdot f_{Y_i=y_i|\mathcal{F}_{i-1}^*}(y) \\ &= f_{X_i|Y_i=y_i,\mathcal{F}_{i-1}^*}(x) \cdot ([\Lambda_i] \cdot [1 - \Lambda_i]) \\ &= \underbrace{f_{X_i|Y_i=y_i,\mathcal{F}_{i-1}^*}(x)}_{f_{GGamma}} \cdot \underbrace{P\left(Y_i = y|\mathcal{F}_{i-1}^*\right)}_{f_{Logistic}}. \end{aligned}$$

3 Estimation and Inference

It is well-known that financial markets pass through hectic periods of increased activity as well as calm slowdowns, reflecting different degrees of liquidity of the asset. Former studies have found a persisting diurnal pattern of trading activities over the course of a trading day, as shown in figure 3, due to the institutional characteristics of organized financial markets (like predetermined opening and closing hours or intraday auctions). As the rate of information arrival will also vary over the trading day, one has to pay regard to this regular daily seasonality. Therefore, smoothing techniques are required to get deseasonalized observations. In this paper a nonlinear kernel regression with a bandwidth h_T of 10 minutes was performed. It

was assumed that the diurnal seasonal component can be computed by the Nadaraya-Watson estimator. Let $\tilde{X}_i$ denote the observed duration, then

$$X_i := \frac{\tilde{X}_i}{m\left(t_i\right)},$$

with

$$\begin{aligned} m\left(t_i\right) &= E\left(\tilde{X}_i | t_i\right) \\ &= \frac{\sum_{i=1}^{n} \tilde{x}_i \cdot K\left(\frac{t-t_i}{h_T}\right)}{\sum_{i=1}^{n} K\left(\frac{t-t_i}{h_T}\right)} \end{aligned}$$

and $K\left(.\right)$ = Gaussian kernel function, is the deseasonalized duration.

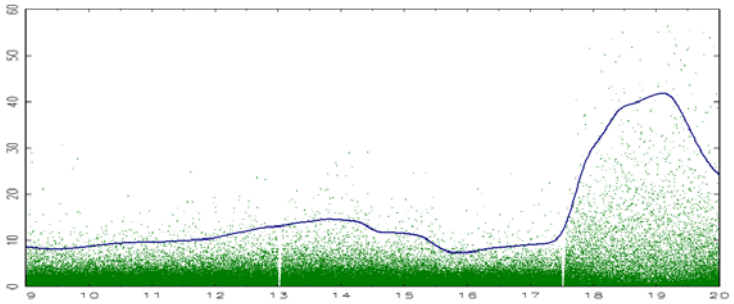

Figure 3: Diurnal pattern

To estimate the bivariate model, one must maximize the two likelihood functions jointly. To ensure the stationarity of the process, one must take care of the very restrictive constraints for the ACD part, in general

$$\begin{aligned} \omega &> 0 \\ \alpha_j, \beta_k, \delta &\geq 0 \qquad \forall j, k \\ \sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k + \delta &< 1 \end{aligned}$$

or, in this study

$$\begin{aligned} \omega' &> 0 \\ \alpha_1, \alpha_2, \beta, \delta, \alpha', \beta' &\geq 0 \\ \alpha_1 + \alpha_2 + \beta + \delta &< 1. \end{aligned}$$

The likelihood of the Generalized Gamma-ACD part is

$$\begin{aligned} L_{ACD} &= L\left(x_1,...,x_n;\boldsymbol{\theta}_1\right) \\ &= \prod_{i=1}^{n} f_{GGamma}\left(x_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_1\right) \\ &= \prod_{i=1}^{n} \frac{\gamma}{x_i\cdot\Gamma\left(\lambda\right)}\left(\frac{x_i}{\Psi_i^{*}}\cdot\frac{\Gamma\left(\lambda+\frac{1}{\gamma}\right)}{\Gamma\left(\lambda\right)}\right)^{\gamma\lambda}\cdot\exp\left\{-\left(\frac{x_i}{\Psi_i^{*}}\cdot\frac{\Gamma\left(\lambda+\frac{1}{\gamma}\right)}{\Gamma\left(\lambda\right)}\right)^{\gamma}\right\} \end{aligned}$$

from which one can derive the log-likelihood

$$\begin{aligned} \mathcal{L}_{ACD} &= \ln L\left(x_1,...,x_n;\boldsymbol{\theta}_1\right) \\ &= \sum_{i=1}^{n}\ln\left(f_{GGamma}\left(x_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_1\right)\right) \\ &= \sum_{i=1}^{n}\ln\left(\frac{\gamma}{x_i\cdot\Gamma\left(\lambda\right)}\right)+\gamma\lambda\cdot\ln\left(\frac{x_i}{\Psi_i^{*}}\cdot\frac{\Gamma\left(\lambda+\frac{1}{\gamma}\right)}{\Gamma\left(\lambda\right)}\right)-\left(\frac{x_i}{\Psi_i^{*}}\cdot\frac{\Gamma\left(\lambda+\frac{1}{\gamma}\right)}{\Gamma\left(\lambda\right)}\right)^{\gamma} \end{aligned}$$

where λ and γ are the specific parameters of the density function (both included in $\boldsymbol{\theta}_1$) and

$$\Psi_i^{*}=\omega'+\sum_{j=1}^{p}\alpha_j X_{i-j}+\sum_{k=1}^{q}\beta_k\Psi_{i-k}+\delta\Lambda_{i-1}.$$

Further, the likelihood-funktion of the dynamic logit model is

$$\begin{aligned} L_{ACL} &= L\left(y_1,...,y_n;\boldsymbol{\theta}_2\right) \\ &= \prod_{i=1}^{n} f_{Logistic}\left(y_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_2\right) \\ &= \prod_{i=1}^{n}\left[\Lambda_i\right]^{y_i}\cdot\left[1-\Lambda_i\right]^{(1-y_i)} \end{aligned}$$

or in the logarithmic form

$$\begin{aligned} \mathcal{L}_{ACL} &= \ln L\left(y_1,...,y_n;\boldsymbol{\theta}_2\right) \\ &= \sum_{i=1}^{n}\ln\left(f_{Logistic}\left(y_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_2\right)\right) \\ &= \sum_{i=1}^{n} y_i\cdot\ln\left[\Lambda_i\right]+\left(1-y_i\right)\cdot\ln\left[1-\Lambda_i\right] \end{aligned}$$

with

$$\Lambda_i=\left[1+\exp\left(-\left(\sum_{j=1}^{u}\alpha_j' Y_{i-j}+\sum_{k=1}^{v}\beta_k'\Lambda_{i-k}+\sum_{l=1}^{r}\sum_{w=1}^{m}\tau_w Z_{w,i-l}\right)\right)\right]^{-1}.$$

In general, the likelihood function of the bivariate model is

$$\begin{aligned} L_{BIV} &= L\left(\left(x_1,y_1\right),...,\left(x_n,y_n\right);\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right) \\ &= \prod_{i=1}^{n} f_{GGamma}\left(x_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_1\right)\cdot f_{Logistic}\left(y_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_2\right). \end{aligned}$$

Taking the logarithm, one gets

$$\begin{aligned} \mathcal{L}_{BIV} &= \ln L\left(\left(x_1,y_1\right),...,\left(x_n,y_n\right);\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right) \\ &= \underbrace{\sum_{i=1}^{n}\ln\left(f_{Logistic}\left(y_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_2\right)\right)}_{\mathcal{L}_{ACL}} + \\ &\quad \underbrace{\sum_{i=1}^{n}\ln\left(f_{GGamma}\left(x_i|\mathcal{F}_{i-1}^{*};\boldsymbol{\theta}_1\right)\right)}_{\mathcal{L}_{ACD}}. \end{aligned}$$

Referring to the $ACD(2,1)\times ACD(1,1)$ model presented in section 3, the functions to be maximized jointly are

$$\begin{aligned} \mathcal{L}_{ACL} &= \sum_{i=1}^{n} y_i\cdot\ln\underbrace{\left[\frac{1}{1+\exp\left(-\left(\alpha_1'Y_{i-1}+\beta_1'\Lambda_{i-1}+\mathbf{Z}_{i-1}^{*}\boldsymbol{\tau}\right)\right)}\right]}_{=\Lambda_i} \\ &\quad +\left(1-y_i\right)\cdot\ln\underbrace{\left[\frac{1}{1+\exp\left(\alpha_1'Y_{i-1}+\beta_1'\Lambda_{i-1}+\mathbf{Z}_{i-1}^{*}\boldsymbol{\tau}\right)}\right]}_{=1-\Lambda_i} \end{aligned}$$

and

$$\mathcal{L}_{ACD}=\sum_{i=1}^{n}\ln\left(\frac{\gamma}{x_i\cdot\Gamma\left(\lambda\right)}\right)+\gamma\lambda\cdot\ln\left(\frac{x_i}{\Psi_i^{*}}\cdot\frac{\Gamma\left(\lambda+\frac{1}{\gamma}\right)}{\Gamma\left(\lambda\right)}\right)-\left(\frac{x_i}{\Psi_i^{*}}\cdot\frac{\Gamma\left(\lambda+\frac{1}{\gamma}\right)}{\Gamma\left(\lambda\right)}\right)^{\gamma}$$

with

$$\Psi_i^{*}=\omega'+\alpha_1X_{i-1}+\alpha_2X_{i-2}+\beta\Psi_{i-1}^{*}+\delta\Lambda_{i-1}$$

and

$$\mathbf{Z}_{i-1}^{*}\boldsymbol{\tau}=\tau_1C_{i-1}^{Ask}+\tau_2C_{i-1}^{Bid}+\tau_3Dur_{i-1}^{Ask}+\tau_4Dur_{i-1}^{Bid}.$$

4 Dataset and Empirical Results

The dataset is extracted from the order book of the German XETRA system for the Deutsche Telekom stocks. The sample includes 143514 observations

from 31st July until 1st September 2000, in total 25 trading days in 5 weeks. The daily trading hours were from 9 a.m. to 8 p.m., interrupted by (at least) two intraday auctions at 1.p.m and 5 p.m. (each lasting at most 120 seconds), as visible in figure 3. The Xetra dataset allows the reconstruction of all quotes and resulting trades, for it contains detailed information on time stamped transaction data. They not only indicate whether the trade is buy or sell initiated but also give almost complete information about the volume of the bid and ask orders. The particular time stamp of bid/ask-orders is also available. (A quote consists of four numbers, a bid and an ask price, and a bid and an ask quantity, called the quoted depth.) The CML-procedure of the Aptech software GAUSS 5.0 was used for **joint** estimation of the model (1) and (2).

An understanding of the time varying speed of transactions is important in practice in order to determine **when** to enter the trading platform to exhaust the temporarily existing consumer/producer surplus in the market. This is possible due to different pricing strategies of asymmetrically informed traders. According to the market microstructure theory, uninformed market participants deduce information in the market from the trading process. Thus, the trading process (not ordering) itself serves as a source of information, and necessarily traders take part to update their knowledge. Knowing the news means reducing the risk. So interesting economic questions for traders are: When will the next event happen? What value should we expect for the mark at the next arrival time? The more information there is in the market the faster they have to react. Estimating the bivariate model proposed above, the ACD part (1) shows the following results:

$$\begin{aligned} E\left(X_i|\mathcal{F}^*_{i-1}\right) &\equiv \Psi^*_i \\ &= \omega' + \alpha_1 X_{i-1} + \alpha_2 X_{i-2} + \beta\Psi^*_{i-1} + \delta\Lambda_{i-1} \end{aligned}$$

with

	$\boldsymbol{\theta}_{ACD}$
ω'	0.0061139728
α_1	0.1479250780
α_2	−0.0782757184
β	0.9219252455
δ	0.0084253949

Empirically, we can clearly see the time dependent ("inter-order") duration cluster, which means that there is a noticeable pattern in order book data: long durations tend to be followed by long durations and short durations tend to be followed by short durations, discernible in figure 4. But the ACD part of the model only describes the temporal distances between events, without distinguishing between different kinds of events.

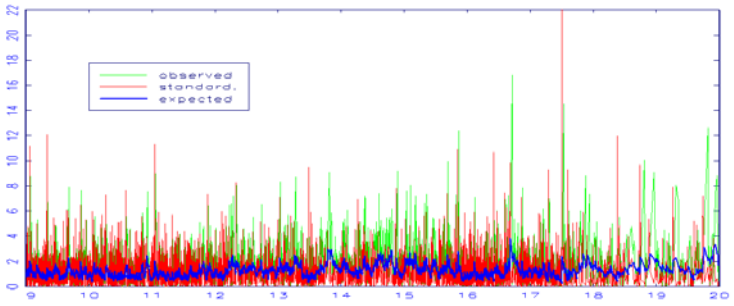

Figure 4: Estimated durations

We must have in mind that only different activities of different market participants make trade possible (one sells, one buys) and are the major source of volatilities. The next aim is to investigate the market side from which we observe activities. New questions would be: When will the next event happen and 'where'? How long should we expect to wait for a particular type of event to occur? What value should we expect for the mark influenced by which market side? As Engle and Lunde (2003) found out, quotes and trades tend to cluster in time in both a deterministic and stochastic way. In this paper, the results of the ACL part (2) of the model show that

$$\begin{aligned} P\left(Y_i = 1 | \mathcal{F}^*_{i-1}\right) &= \Lambda(\alpha_1' Y_{i-1} + \beta_1' \Lambda_{i-1} \\ &\quad + \tau_1 C^{Ask}_{i-1} + \tau_2 C^{Bid}_{i-1} + \tau_3 Dur^{Ask}_{i-1} + \tau_4 Dur^{Bid}_{i-1}) \end{aligned}$$

with

	$\boldsymbol{\theta}_{ACL}$
α_1'	0.1763916229
β_2'	0.4681478040
τ_1	0.0242570517
τ_2	−0.0333822391
τ_3	0.0070851030
τ_4	0.0104013503

The probability of a bid order will become larger, if the last order type was a bid , $\alpha' > 0$. Further, the bid order probability will be the larger, (a) the larger the preceding bid order probability, $\beta' > 0$, (b) the longer the ask queue, $\tau_1 > 0$, (c) the shorter the bid queue, $\tau_2 < 0$, and (d) the longer ago the last bid/ask order.

According to the theoretical findings, the total volume in the particular queues characterizes the demand and supply side. The traded quantities on the particular sides of the market are strong proxies for the existence of

information at the current time. Usually the difference between the transaction price and the current midquote characterizes the depth associated with the last transaction. The higher this difference is, the more volume is absorbed from the particular queue (order type cluster) which should decrease the probability of the occurrence of a trade of the same type in the next instant (no cluster). All estimates and their standard errors are reported in the following table:

sample size =	143514.000	$E(\varepsilon_i) =$	1.0002528782
likelihood =	-237600.5829912351		
parameter	**coefficient**	**std. error**	**t-value**
$\boldsymbol{\theta}_{ACD} = \boldsymbol{\theta}_1$			
ω'	0.0061139728	0.0005902272	10.3586765997
α_1	0.1479250780	0.0041761731	35.4212036369
α_2	−0.0782757184	0.0054561849	−14.3462364020
β	0.9219252455	0.0026782565	344.2259029341
δ'	0.0084253949	0.0004938009	96.4060767646
f_{GGamma}			
γ	0.6003136836	0.0062269278	54.4953617999
λ	2.1312924787	0.0391096124	17.0623309274
$\boldsymbol{\theta}_{ACL} = \boldsymbol{\theta}_2$			
α'_1	0.1763916229	0.0119725800	14.7329666874
β'_2	0.4681478040	0.0068992552	67.8548324923
τ_1	0.0242570517	0.0008502534	28.5292034568
τ_2	−0.0333822391	0.0011153979	−29.9285464747
τ_3	0.0070851030	0.0003137292	22.5834984107
τ_4	0.0104013503	0.0002278614	45.6477016957

5 Conclusion

This paper develops a bivariate modelling framework for analyzing the arrival process of ask and bid orders in an electronic order book market. The econometric approach consists of two parts: In the first step, a simple logit model is run in order to analyze the determinants of the order type´s transition conditioned on the past durations and the last state of the order book. In the second step, in order to recover the whole temporal structure of all time stamped events, the common ACD model is extended by affixing the dynamic logit model and additional covariates. The main idea is to base both conditional functions on two components jointly, one to model duration clusters, one to describe the order type with a time dependent probability function revealing the information flow and trading activity of the market.

Using detailed transaction data from the German XETRA system, a few new counting variables are generated as further time varying covariates reflecting the state of the order book. The empirical results show that characteristics associated with previous orders and durations as well as the last state of the order book have a significant impact on the traders´ decisions when to trade and on which side of the market. Obviously, traders pay strong attention to the order arrival process and the corresponding queues of the order book. The inclusion of the dynamic logit component substantially improves the fit of the original ACD model, providing deeper insights into the joint market dynamics.

References

[1] **Amemiya, Takeshi (1985):** *Advanced Econometrics*, Basil Blackwell, Oxford.

[2] **Alexander, Carol (2001):** *Market Models - A Guide to Financial Data Analysis*, John Wiley & Sons, Chicester et al.

[3] **Andersen, Erling B. (1997):** *Introduction to the Statistical Analysis of Categorial Data*, Springer, Berlin et al.

[4] **Baltagi, Badi H. (2001):***Econometric Analysis of Panel Data*, John Wiley & Sons, England.

[5] **Bauwens, Luc et al. (2000):** *A Comparison of Financial Duration Models via Density Forecast*, in: Econometric Society World Congress 2000 Contributed Papers, No. 810.

[6] **Bauwens, Luc / Giot, Pierre (2001):** *Econometric Modelling of Stock Market Intraday Activity*, Kluwer Academic Publishers, Dordrecht, NL.

[7] **Bauwens, Luc / Hautsch, Nikolaus (2003):** *Stochastic Conditional Intensity Process*, Working Paper, CORE & CoFE.

[8] **Blossfeld, Hans-Peter / Hamerle, Alfred / Mayer, Karl U. (1989):** *Event History Analysis*, Lawrence Erlbaum Associates Publishers, New Jersey.

[9] **Bowsher, Cliff (2002):** *Modelling Security Market Events in Continuous Time: Intensity Based, Multivariate Point Process Models*, Nuffield Economics Discussion Paper Series, University of Oxford.

[10] **Cameron, A. Colin / Trivedi, Pravin K. (1998):** *Regression Analysis of Count Data*, Cambridge University Press, UK.

[11] **Campbell, John Y. / Lo, Andrew W. / MacKinley, A. Craig (1997):** *The Econometrics of Financial Markets*, Princeton University Press, Princeton.

[12] **Dacorogna, M. et al. (2001):** *An Intorduction To High-Frequency Finance*, Academic Press, San Diego.

[13] **Drost, Feike C. / Werker, Bas J.M. (2001):** *Semiparametric Duration Models*, Working Paper, Tilburg University, No. 2001-11.

[14] **Dufour, Alfonso / Engle, Robert F. (2000):** *The ACD-Modell: Predictability of the Time Between Consecutive Trades*, unpublished, downloaded from JEL.

[15] **Engle, Robert F. (2000):** *The Econometrics of Ultra-High-Frequency Data*, in: Econometrica, Vol. 68, No. 1, P. 1-22.

[16] **Engle, Robert F. / Lunde, Asger (2003):** *Trades and Ouotes: A Bivariate Process*, in: Journal of Financial Econometrics, Vol. 1, No. 2, P. 159-188.

[17] **Engle, Robert F. / Russell, Jeffrey R. (1997):** *Forecasting the frequency of changes in quoted foreign exchage prices with the autoregressive conditional duration model*, in: Journal of Empirical Finance, Vol. 4, P. 187-212.

[18] **Engle, Robert F. / Russell, Jeffrey R. (1998):** *Autoregressive Conditional Duration: A New Model for Irregularly-Spaced Financial Transactions Data*, in: Econometrica, Vol. 66 , P. 1127-1162.

[19] **Engle, Robert F. / Russell, Jeffrey R. (2002):** *Econometric Analysis of Discrete-valued Irregularly-Spaced Financial Transactions Data*, Working Paper, University of Chicago & University of California.

[20] **Fernandes, Marcelo / Grammig, Joachim (2000):** *A Family of Autoregressive Conditional Durations Models*, Working Paper 2001/36, CORE.

[21] **Gourieroux, Chiristian / Jasiak, John (2001):** *Financial Econometrics*, Princeton University Press, Princeton.

[22] **Greene, William H. (2003):** *Econometric Analysis*, 5th ed., Prentice Hall, New Jersey.

[23] **Hall, Anthony / Hautsch, Nikolaus / MacCulloch, James (2003):** *Estimating the Intensity of Buy and Sell Arrivals in a Limit Order Book Market*, Working Paper, CoFE & UTS.

[24] **Hautsch, Nikolaus (2002):** *Testing the Conditional Mean Function of Autoregressive Conditional Duration Models*, Working Paper, CoFE.

[25] **Johnson, Norman L. / Kotz, Samuel / Balakrishnan, N. (1994):** *Continuous univariate Distributions*, Vol.1, 2nd ed. , John Wiley & Sons, New York.

[26] **Johnson, Norman L. / Kotz, Samuel / Balakrishnan, N. (1995):** *Continuous univariate Distributions*, Vol.2, 2nd ed. , John Wiley & Sons, New York.

[27] **Judge, George G. / Miller, Douglas J. / Mittelhammer, Ron C. (2000):** *Econometric Foundations*, Cambridge University Press, UK.

[28] **Lancaster, Tony (1990):** *The Econometric Analysis of Transition Data*, Cambridge University Press, New York et al.

[29] **Liesenfeld, Roman / Pohlmeier, Winfried (2003):** *A Dynamic Integer Count Data Model for Financial Transaction Prices*, Discussion Paper Feb. 2003, CoFE.

[30] **O´Hara, Maureen (1997):** *Market Microstructure Theory*, Blackwell, Oxford.

[31] **Russell, Jeffrey R. (1999):** *Econometric Modelling of Multivariate Irregularly-Spaced High Frequency Data*, unpublished, downloaded from JEL.

[32] **Winkelmann, Rainer (1998):** *Econometric Analysis of Count Data*, 2nd ed., Springer, Heidelberg.

[33] **Wooldridge, Jeffrey M. (2002):** *Econometric Analysis of Cross Cection and Panel Data*, MIT Press, Massachusetts.