INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2001-19

On the (Intradaily) Seasonality and Dynamics of a Financial Point Process : A Semiparametric Approach

D. VEREDAS[1]
J. RODRIGUEZ-POO[2]
A. ESPASA[3]

[1] CORE, Université Catholique de Louvain, 34 Voie du Roman Pays, B1348, Louvain-la-Neuve, Belgium. Email : veredas@core.ucl.ac.be
[2] Departamento de Economia, Universidad de Cantabria. Email : rodrigjm@unican.es
[3] Statistics and Econometrics Department, Universidad Carlos III de Madrid. Email : espasa@econ-est.uc3m.es

RESUME

Les mouvements saisonniers pendant la journée et la semaine sur les marchés boursiers sont idiosyncrasiques et doivent donc être pris en compte dans la modélisation des données. Dans ce papier, nous proposons un modèle à composantes pour l'analyse des durées financières. Le modèle est semiparamétrique, la saisonnalité étant calculé de façon nonparamétrique et la dynamique étant spécifiée paramétriquement. L'estimation est jointe, les paramtres et la courbe de saisonnalité sont estimés de façon consistant et efficace. En outre, nous montrons que les durées nulles sont informatives et une façon de les traiter est proposée. La méthode est appliquée au processus des durées inter-transaction de Bankinter, une banque espagnole moyenne cotée sur la Bolsa de Madrid.

ABSTRACT

Seasonal patterns during the day and during the week in the stock exchange markets are idiosyncratic and thus they should be taken into account when modelling these data. In this paper we propose a component model for the analysis of financial durations, that can be extended easily to any other high frequency financial variable. The model is semiparametric where the seasonality is computed nonparametrically and the dynamics are specified parametrically. Estimation is joint and parameters and seasonal curve are proved to be consistent and efficient. Additionally, durations equal to zero are shown to be informative and a way to deal with them is proposed. All the methodology is applied to the trade duration process of Bankinter, a medium size spanish bank traded in Bolsa de Madrid

Keywords: Tick-by-tick, ACD model, Seasonal analysis, Nonparametric methods.

JEL classification: C14, C41, G10.

[0]We thank Luc Bauwens, Joachim Grammig and Jorge Yzaguirre for useful remarks. We also thank Joachim for lending us his computer for "some" time. First author acknowledges to Christian Gouriéroux for his support at CREST, since a part of the paper were done while the first author was in CREST as an EDP student. He also acknowledges financial support from the Université catholique de Louvain (project 11000131). The second author acknowledges financial support, while writting this paper, from the Institute of Statistics at the Université catholique de Louvain. Finally third author acknowledges financial support from the Spanish Ministry of Education (project PB98-0140). The scientific responsibility is assumed by the authors.

1 Introduction

The issue of modeling financial duration processes is a fashionable area of research since Engle and Russell (1998) introduced the Autoregressive Conditional Duration (ACD) model. Its analysis is justified from an economic and a statistical point of view. On one hand, market microstructure theory shows that time between events in a stock exchange market conveys information and thus time has to be analyzed. On the other hand, since data are the "so called" tick-by-tick data,[1] they are nothig else than a one dimensional point process, with time as space. Thus time is the random variable of the point process and in each point there is an associated vector of marks, and both time and the marks can be modeled.

Since the former model a plethora of modifications and alternatives have been proposed. Among others, Bauwens and Giot (2000) introduced the Log-ACD model, which is an exponential version of the ACD. Grammig and Mauer (2000) used a Burr distribution in the ACD model. Zhang et al. (1999) introduced a threshold ACD. Drost and Werker (2001) provide a method to obtain efficient estimators of the ACD model without need to specify the distribution. Camacho and Veredas (2001) consider the analysis of a bivariate duration process using random aggregation techniques. Alternative models are the the Stochastic Conditional Duration (SCD) model of Bauwens and Veredas (1999) and the Stochastic Volatility Duration (SVD) model of Ghysels et al. (1998) which are both based on latent factor models. Almost all these models are surveyed in Bauwens et al. (2000).

In most of the above studies, the variable that has been considered show a strong intradaily and intraweekly seasonality. In a explanatory graphic analysis the strong seasonal component is detected by the presence of the U (or inverted U) shape that ultra high frequency financial variables exhibit during the day and during the week (see figure 5 in section four).

This problem is well known when dealing with regularly spaced variables, that is, when dealing with variables that are observed every fixed periods of time. Moreover this analysis has focused mainly in the volatility's intradaily behaviour of either an stock exchange market or a foreign exchange (FX) market. Engle et al. (1990) analyze how the information flow is transmitted through world regions in the FX market using hourly data. Harris (1986) does a panel data analysis using 15 minute interval returns data of firms traded in NYSE. Baillie and Bollerslev (1990) studied the intra-day and inter-market FX volatility using a qualitative approach with hourly data. Bollerslev and Domowitz (1993) do a similar analysis but for returns and bid-ask spread of the deutsche mark-dollar exchange rate using data recorded at 5 minute intervals. Andersen and Bollerslev (1997) used a frequency domain approach for filtering the five minutes deutsche mark-dollar exchange rate and getting rid off the seasonal pattern. Andersen and Bollerslev (1998) used a

[1] Other ways to term these type of data are high frequency data (HFD) or ultra high frequency data (UHFD). We prefer either tick-by-tick or point process, since using the former two implies that we assume the existence of a certain frequency and one of the main characteristics of these kind of processes is the lack of periodicity.

diffent approach and analyze the intradaily and intraweekly seasonality using spectral analysis and they took into account macroeconomic announcements. Finally, Beltratti and Morana (1999) used half hour deutsche mark-dollar exchange rate and they modeled it following a structural approach "*à* la Harvey".

All these previous works have been done using regularly spaced data (hourly, half-hourly, 15 minutes, or 5 minutes). In tick-by-tick data the most popular approach for dealing with intradaily seasonality was introduced by Engle and Russell (1998). The method consists in estimating the intradaily seasonality by means of a piecewise cubic spline. Although Engle and Russell (1998) apparently succeed in the joint estimation of the parameters of the cubic spline and the ACD model, it is a hard task and the convergence towards a global maximun is not assured. Because these reasons most of other studies have focused in a two step procedure, where in the first step, the inverted U shape is removed through some filter and, in a second step, the ACD model is estimated by using the deseasonalized variables. The filter basically consists in calculating the average durations every, say, 30 minutes and then smooth this piecewise constant function through cubic splines. Alternatively Gouriéroux et al. (1999) analyzed the intraday market activity using kernels for the intraday intensity as well as for the survivor function, but they do not differentiate between seasonal pattern and long-run dynamics. Gerhard and Haustch (2000) proposed a model for financial durations using a proportional hazard model where seasonality is modeled using a flexible Fourier transform.

The two step procedure presents some serious drawbacks. Mainly it performs accurately if both the seasonal and the non-seasonal components depends on some deterministic time index, and the non-seasonal dynamics of the duration process is linear in the parameters to be estimated. Otherwise, the two step estimation procedure can lead to serious misspecification errors.

In this paper we assume that tick-by-tick processes can be decomposed in two components that stands for the short-run and the long-run. The short-run refers to the intradaily and intraweekly seasonality while the long-run can be considered as the core dynamics of the process.

In the standard theory of time series, two approaches exist for dealing with these components. The first one considers that any time series can be analyzed by means of an ARMA model that, using different lags in the polynomials and exogenous variables, account for the components. The second approach assumes that the time series can be decomposed in latent components which are not observed but have some dynamics and/or some patterns.

In the framework of tick-by-tick data, the ARMA approach is not feasible since one of the main characteristics of these data is the lack of periodicity. Therefore we focus in the second approach, assuming the decomposition of the time series in components that are estimated separately but not independently. In order to do so, we rely on the assumption that the conditional expectation of the duration of some financial variable can be decomposed in the three mentioned terms. Under this assumption, they can be estimated simultaneously.

The short-run component is modeled nonparametrically and the long-run component is assumed to belong to the parametric ACD family. Both components are estimated simultaneously by maximizing alternatively a local and a global version of the likelihood function respectively. Under the correct choice of the smoothing parameter, this estimation method provides root-N consistent semiparametric estimators of the parameters of the Log-ACD model. Furthermore, if the conditional likelihood is correctly specified the estimators are efficient.

We also deal with durations equal to zero. These durations are often found in the trade process. Previous studies eliminate them using the microstructure argument that all the trades executed in the same second come from the same trader that has split a big order block in small blocks. We show that this is not always true and, indeed, most of the times the durations zero are clustered around round prices due to the fact that the limit orders of the retail traders are set for being executed at the round prices and hence trades executed in the same second do not belong to the same trader but to many retail traders.

The plan of the paper is a as follows. Section two develops a general framework for analyzing tick-by-tick financial variables, decomposing the process in the two above mentioned terms. Notice that even if notation and empirical application is done for duration processes, any other variable can be easily analyzed. Section three is devoted to the analysis of each one of components introducing a modelling strategy for the analysis of durations equal to zero and a discussion on how to model the seasonal component. Section four estudies the theoretical properties of the estimators, that is, the parameters and the non parametric curve, proving consistency and normality for the parameters and the best estimator for the curve. Finally section five is devoted to the empirical application, doing an analysis to the trading duration process of Bankinter, a medium size Spanish bank traded in Bolsa de Madrid.

2 Basic Econometric Model

In order to introduce the main contribution of our paper, we need to establish a basic econometric framework. Following Engle and Rusell (1998) and Engle (2000), let t_i be the time at which the i-th trade occurs and let $d_i = t_i - t_{i-1}$ be the duration between trades. Let us consider also that we have observed k marks at the i-th event, y_i. Then, we have available the following set of observations

$$\{(d_i, y_i)\}_{i=1,\cdots,n}.$$

Furthermore, assume that the i-th observation has the joint density conditional on the past filtrations as

$$(d_i, y_i) | \, I_{i-1} \sim f\left(d_i, y_i | \bar{d}_{i-1}, \bar{y}_{i-1}; \delta\right),$$

where $\bar{z}_i = (z_i, z_{i-1}, \cdots, z_1)$ is the present and past information of the z stochastic process and δ is a set of parameters in some possibly infinite dimensional space.

Within this statistical framework, our aim is to estimate this parameter vector δ (or any nonlinear combination of its components) by using maximum likelihood techniques. To this end, we construct the following likelihood function

$$L_n\left(d, y; \delta\right) = \sum_{i=1}^{n} \log\, f\left(d_i, y_i | \bar{d}_{i-1}, \bar{y}_{i-1}; \delta\right). \tag{1}$$

Following a reduction process we can considerably simplify the previous log-likelihood expression. Without loss of generality we can write

$$\log\, f\left(d_i, y_i | \bar{d}_{i-1}, \bar{y}_{i-1}; \delta\right) = \log\, p\left(d_i | \bar{d}_{i-1}, \bar{y}_{i-1}; \delta_1\right) + \log\, g\left(y_i | \bar{d}_i, \bar{y}_{i-1}; \delta_2\right),$$

where $\delta = (\delta_1, \delta_2)$. Moreover, if both the parameter vector δ are variation free and the marks, y, are defined as weakly exogenous for the parameters of interest δ_1, then the maximization of (1) is equivalent to the maximization of the following likelihood function

$$L_n\left(d, y; \delta_1\right) = \sum_{i=1}^{n} \log\, p\left(d_i | \bar{d}_{i-1}, \bar{y}_{i-1}; \delta_1\right). \tag{2}$$

The exogeneity assumption is crucial and arguable. For example, if y_i is the volatility of the tick-by-tick process, we are assuming that there exists an unidirecccional causality, i.e. volatility causes the durations but not the contrary. This relationship has been pointed out by Ghysels (2000), among others, and in terms of market microstructure it seems that a joint analysis of (d_i, y_i) is more adequate. However, it is out of the scope of this paper and we let this issue for further research. Thus, if the conditional density is correctly specified, then standard maximum likelihood techniques apply and the maximum likelihood estimator of δ_1 is consistent and asymptotically normal. Alternatively, as pointed out in Engle and Rusell (1998) and Engle (2000) it would be of interest to have available some estimation techniques that do not require the knowledge of the conditional density function. Two alternative approaches that allow for consistent estimation of the parameters of interest without specifying the conditional density are Quasi Maximum likelihood techniques, QML, (see Gouriéroux, Monfort and Trognon, 1984) and Generalized Linear Models, GLM, (see McCullagh and Nelder, 1989). In both approaches, it is assumed that the duration variable d, conditonally on past values of d and y depends on a scalar parameter $\theta = h\left(\bar{d}_{i-1}, \bar{y}_{i-1}; \delta_1\right)$, and its distribution forms a one dimensional exponential family with conditional density

$$p\left(d_i | \bar{d}_{i-1}, \bar{y}_{i-1}; \theta\right) = \exp\left(d_i\theta - b(\theta) + c(d_i)\right),$$

where $b(\cdot)$ and $c(\cdot)$ are some known functions. The main difference between the QML and the GLM approach is simply a different parametrization of this exponential family. Here in this paper we will adopt for convenience the GLM approach. Then it is straightforward to see that the Maximum Likelihood estimator of θ

solves the following first order conditions: $\sum_j \{d_j - b'(\theta)\} = 0$. Furthermore, since by the properties of the exponential functions,

$$E\left[d_i|\bar{d}_{i-1}, \bar{y}_{i-1}\right] = b'(\theta) = \mu\left\{\bar{d}_{i-1}, \bar{y}_{i-1}; \delta_1\right\} \tag{3}$$

and

$$\text{Var}\left[d_i|\bar{d}_{i-1}, \bar{y}_{i-1}\right] = b''(\theta) = \sigma^2 V\left\{\mu(\bar{d}_{i-1}, \bar{y}_{i-1}; \delta_1)\right\}, \tag{4}$$

then the M.L.E. estimator of θ can also be obtained from the solution to the following equation

$$\sum_{i=1}^{n} \frac{(d_i - \mu(\theta))\,\mu'(\theta)}{V(\mu(\theta))} = 0. \tag{5}$$

As it can be clearly realized from equations (3), (4) and (5) the estimation of the parameter of interest θ (the so called canonical parameter) can be performed without needing to specify the whole conditional distribution function. It is only necessary to specify the functional form of the conditional mean, $\mu(\cdot)$, and the conditional variance $V(\cdot)$, but not the whole distribution. Engle and Rusell (1998) propose to specify the conditional mean function by using the ACD class of models that consists on parametrizations such as

$$E\left[d_i|\bar{d}_{i-1}, \bar{y}_{i-1}\right] = \mu\left(\bar{d}_{i-1}, \bar{y}_{i-1}; \delta_1\right) = \varphi\left(\omega + \sum_{j=1}^{J} \alpha_j g(d_{i-j}) + \sum_{k=1}^{K} \beta_k \mu_{i-k}\right), \tag{6}$$

where the parameters of interest would be $\delta_1 = (\omega, \alpha_1, \cdots, \alpha_J, \beta_1, \cdots, \beta_K)$. The functions $\varphi(\cdot)$ and $g(\cdot)$ take the values $\varphi(s) = s$ and $g(s) = s$ for the ACD model and $\varphi(s) = \exp(s)$ and $g(s) = \ln(s)$ for the Log-ACD model. The relationship between the predictors in equation (6) and the canonical parameter is given by the so called *link function*. This function is going to depend on the member of the exponential family that we are going to use. For the exponential distribution the link function is

$$\theta = -\frac{1}{\varphi\left(\omega + \sum_{j=1}^{J} \alpha_j g(d_{i-j}) + \sum_{k=1}^{K} \beta_k \mu_{i-k}\right)}. \tag{7}$$

Noting that under this distribution $\mu(\theta) = -\theta^{-1}$ and $V(\mu(\theta)) = \mu^2$, then (5) are the first order conditions for the maximization of the log-likelihood function for exponentially distributed data.

As it has been pointed out in many recent studies, the ACD specification is sometimes too simple since the expected duration can vary over time, or can be subject to many different time effects. One way to extend the previous model is to decompose the conditional mean in different effects. In the standard time series literature any stochastic process can be decomposed in a combination (we adopt a multiplicative decomposition being the additive straightforward) of cycle and trend, seasonal pattern and noise, i.e. $X_t = X_t^{CT} \cdot S_t \cdot \varepsilon_t$. This decomposition, of long tradition in time series analysis, has been already used in volatility analysis

(see for example Andersen and Bollerslev, 1998). In ultra high frequency data, the ACD model has been usually estimated using a duration time series that was already adjusted by seasonality by using averages over some period of time and piecewise cubic splines to smooth these averages. More precisely, let us denote by d_i the duration variable, let d_i^a be the "diurnally adjusted" duration and finally let $\phi(t_{i-1})$ be the seasonal component. Then the duration is "diurnally adjusted" by

$$d_i^a = \frac{d_i}{\phi\left(t_{i-1}\right)} \tag{8}$$

and the expected duration can be written as

$$E\left[d_i|\bar{d}_{i-1}, \bar{y}_{i-1}\right] = \phi\left(t_{i-1}\right) E\left[d_i^a|\bar{d}_{i-1}, \bar{y}_{i-1}\right]. \tag{9}$$

See, among others, Engle and Russell (1997), and Bauwens and Giot (2000). Note that for (8) and (9) to hold, the time effect must be deterministic. However, as it was argued in the previous section, this assumption rules out possible impacts of factors on seasonality such as macroeconomic announcements, market conditions and other effects. In order to allow for these effects, then it is necessary to assume that both trend-cycle and seasonal components are functions of random variables, and therefore simultaneous estimation of both terms is required. Further, in this paper we propose the following nonlinear structure for the conditional mean

$$E[d_i|\bar{d}_{i-1}, \bar{y}_{i-1}] = \varphi\left(\psi(\bar{d}_{i-1}, \bar{y}_{i-1}; \vartheta_1), \phi(\bar{d}_{i-1}, \bar{y}_{i-1}; \vartheta_2)\right). \tag{10}$$

The function $\varphi(u, v)$ can nest a great variety of models. $\varphi(u, v) = (u \times v)$ stands for an ACD representation whereas $\varphi(u, v) = \exp(u + v)$ represents a Log-ACD representation. Then, following (10) the durations, volatility, trading intensity and volume (in a high frequency framework) can be modelled as a possibly nonlinear function of two components that represents the long-run, $\psi(\cdot; \vartheta_1)$ and the short-run, $\phi(\cdot; \vartheta_2)$, respectively. The long-run component can be considered as the core dynamics and on it the dynamics of the process are modelled. It can be done using autoregressive models (like GARCH or ACD), latent factor models (like SV and SCD) or any other alternative. The short-run component represents the seasonal pattern, that can be intradaily and intraweekly. The next issue is how to specify each of these components. Alternatively another component could be added to (10) accounting for the news effect. Then this third component would be the short-run component because since we are working with tick-by-tick data, short-run means some hours and usually the effect of a news in the stock remains for no more than a couple of hours, as documented by Payne (1996) and Almeida et al. (1996).

3 Specification of the different components

The following natural question is how to model each one of the components. As a first guess, we should chose between a fully nonparametric approach, a semiparametric or a fully parametric. Since we have to specify two different components it

would be sensible to specify parametrically those functions where a lot of information is available, whereas in the case of ignorance a fully nonparametric approach is much more feasible. For the long-run component we adopt some previous prespecified parametric form. The seasonal component is much less investigated, and to our knowledge it does not exist a standard an accepted form for this type of models. On these grounds, we choose to leave it unspecified in the form of a nonparametric function. Furthermore, the interest of the analyst is to predict the process as a whole, that is predict the raw data and not the adjusted one. This is an additional reason for modeling parametrically the component that conveys the past information whilst the deterministic pattern is approached nonparametrically.

We first introduce the specification for the long run component whereas the specification for the short run component will be introduced later in the section.

For the long-run Engle and Russell (1998) introduced the ACD model that accounts for these features. Since this model, more refinated versions has appear in the literature. See Bauwens et al. (2000) for a survey about these kind of models. A version of particular interest is the Log-ACD model of Bauwens and Giot (2000). They model the expected duration exponentially, similarly to the EGARCH model for volatility. This model is useful because it avoids the positivity restrictions of the parameters of the dynamic equation.

A drawback of the ACD and the Log-ACD models, as well as all financial duration models existing in the literature, is that they do not permit durations equal to zero, since the "good" distributions used for durations are not defined at zero.[2] In the exchange markets this a quite common event when dealing with transaction data, where several transaction occurs at the same time.[3] As contrary to other studies, we believe that they convey information since they are the result of limit orders of retail traders posted for being executed at round prices. When dealing with this particular type of durations we are willing of substituting the durations equal to zero for some quantity. This quantity can be either estimated or chosen ad hoc. In any case it should be between zero and one, since the time measure in one second and it is the smallest possible observable duration. This analysis does not only implicates to durations zero but also to the next positive duration after several successive durations zero. That is in order to maintain that the sum of all the durations remains equal to the total time spell considered and, if durations zero are substituted by a certain positive value between zero and one, we should modify the duration strictly positive that occurs after successive durations zero. In terms of time deformation it means that durations zero are enlarged while

[2]The term "good" is because a distribution like the exponential is defined at zero but it is "not good" for financial durations.

[3]This is not entirely true since if we consider time as continuous then by definition there cannot be two events at the same time but, since the minimum time measure is the second, it happens often.

the next strictly positive duration is shrunk. More formally

$$d_i^* = \begin{cases} d_i & \text{if } \ d_i > 0 \text{ and } d_{i-1} > 0 \\ c_i & \text{if } \ d_i = 0 \\ d_i - \sum_{j=1}^{J} c_j & \text{if } \ d_i > 0 \text{ and } d_{i-j} = 0,\ j = 1, \dots J, \end{cases} \tag{11}$$

where J is the number of immediately past successive durations zero. This transformation is subject to the constraints $0 < c_i \leq 1$ and $0 < \sum_{j=1}^{J} c_j \leq 1$. There is one especial case when $d_i = 1$ and $d_{i-1} = 0$. Then d_i is also considered as a duration zero but then if next duration, d_{i+1}, is strictly positive it is not transformed.

Thus, given this transformation, what is of primary interest is to set the values c_i. There exist several alternatives depending on the interest of the analysis. The first approach consists in replacing c_i by some constant ad hoc. The second approach is to estimate them. Estimation can be done considering the model as a left censored model where the censoring is that we do not observe values bellow one. Another possibility would be consider that the duration zero generating process differs from the generation process of the strictly positive durations. Then we can use a similar technique to the hurdle models used in count data.

The principal drawback of these models is that we are dealing with dynamical processes and hence either censoring or hurdle in these processes is not as easy as in the static case since we have to integrate with respect to past censoring and tractability is not assured (see, for example Wei, 1997, for a Bayesian approach to dynamic Tobit models).

The first of the constraints, $0 < c_i \leq 1$, gives us a hint about a possible specification of the parameter c_i. Since it can be time varying and it must be between 0 and 1, one possible functional form is by means of a logistic function and so its value may depend on extra variables such as the number of successive zeros, past durations, prices, etc. Alternatively, any other function that ranges between zero and one can be used. Any distribution function would be valid, specially the cdf's used here, as the Burr or the generalized gamma. These approaches are with no doubt cumbersome and they are themselves subject of a proper research.

Hence, in our framework, since we are mainly interesting in analyzing the intradaily seasonality but without despise the information content in duration zero, we substitute durations zero by $c_j = 1/J$ where J is the number of successive durations zero. The drawback of this approach is that durations zero are considered to be regularly spaced within the second in which they arrive. However, this transformation carries out the above scheme and the constraints are fulfilled. We adopt the easiest approach not expecting great results and letting this subject open for future research.

Following, the specification of the long-run component is done by means of the conditional expectation of a Log-ACD model

$$\log \psi(\bar{d}_{i-1}, \bar{y}_{i-1}; \vartheta_1) = \omega + \alpha \ln d_{i-1} + \beta \psi_{i-1} \tag{12}$$

With respect to the so called short-run component, $\phi(\cdot, \vartheta_2)$, several alternative approaches are available. When modelling seasonality, in this type of models, it

is usually assumed that the seasonal term is somehow related to the time t_i at which the i-th transaction occurs through some smooth function on time. Clearly, it is of great interest to specify this function $\phi(\cdot, \vartheta_2)$. Several proposals have been made in the literature. One might assume that this function belongs to a pre-specified family of parametric functions. This can be the case when using seasonal dummy variables, or truncated trigonometric polynomials. However, in this paper we do not want to specify such a function, and hence we will only assume some smoothness conditions on it. We propose to estimate by using kernel methods that are carefully explained in next section. There exist many reasons to do that: First, they are computationally efficient, second, they allow for easy simultaneous estimation of all parametric and nonparametric components, and finally, we can show the statistical properties of the resulting estimator.

Finally,consider an additional third component, $\rho(\cdot; \varrho)$, for the news effect. It could be expressed as $\rho(\cdot; \varrho) = \sum_{j=1}^{J} \sum_{k=1}^{K} \delta_j (1 - \varrho)^k D_{ijk}$ for $0 < \varrho < 1$ and $\theta_2 = (\delta_1, \cdots, \delta_J, \varrho)$. That is, at a moment i there is a news event that affects to the market up to observation K arrives. The dependency of the variable of interest in the news decreases geometrically since markets do not react instantaneously to the news, but they take some time to dissipate the new information.

4 Simultaneous estimation procedure

Now, once the general model is specified, it is necessary to provide an estimation method that accounts for the unknown quantities that need to be estimated. From equation (10) in Section 2 and the parametric model for the long-run, then we propose the following expression for the mean of the conditional duration model

$$E\left[d_i | \bar{d}_{i-1}, \bar{y}_{i-1}\right] = \varphi\left(\psi(\bar{d}_{i-1}, \bar{y}_{i-1}; \vartheta_1), \phi(t_i)\right),$$

where the function $\psi(\cdot, \vartheta_1)$ is known and the other quantities, ϑ_1 and the function $\phi(\cdot)$ evaluated at time points $t_1, \cdots, t_n$ need to be estimated. This estimation problem is semiparametric since a nonparametric component, ϕ, needs to be estimated jointly with a parametric one ϑ_1. Under this setting standard (quasi-)maximum likelihood techniques do not apply directly and some developments are needed. This extension is based on the so called conditionally parametric approach introduced in Severini and Wong (1992). The basic idea of this method is to estimate the nonparametric function $\phi(\cdot)$ by maximizing a local likelihood function (see Staniswalis, 1989) and simultaneously estimate the parameter vector ϑ_1 by maximizing the un-smoothed likelihood function. If we specify only the conditional mean and the underlying density is assumed to belong to the family of exponential densities, then maximum likelihood methods are available (Severini and Staniswalis, 1994 ; Fan, Heckman and Wand, 1995). Unfortunately, the statistical results from these papers do not apply directly in our case since they assume independent observations.

The (quasi-)likelihood function takes the form

$$Q_n(d,\varphi) = \sum_{i=1}^{n} Q\left(\varphi\left(\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1),\phi(t_i)\right);d_i\right), \tag{13}$$

where the quasi likelihood $Q(\cdot)$ is obtained by integrating (5), i.e.

$$Q(d,g) = \int_g^d \frac{(s-d)}{V(s)} ds.$$

For fixed values of ϑ_1, let us define $\hat{\phi}_{\vartheta_1}(\tau)$ as the solution to the following optimization problem

$$\hat{\phi}_{\vartheta_1}(\tau) = \arg\max_{\eta} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tau - t_i}{h}\right) Q\left(\varphi\left(\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1),\eta\right);d_i\right)$$

for $\tau \in [a,b]$. Then $\hat{\phi}_{\vartheta_1}(\tau)$ must fulfill the following first order conditions

$$\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\tau - t_i}{h}\right) \frac{\partial}{\partial\eta} Q\left(\varphi\left(\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1),\eta\right);d_i\right) = 0. \tag{14}$$

The values of ϑ_1, $\hat{\vartheta}_{1n}$ are obtained as the solution to the following (un-smoothed) optimization problem

$$\hat{\vartheta}_{1n} = \arg\max_{\vartheta_1} \sum_{i=1}^{n} Q\left(\varphi\left(\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1),\hat{\phi}_{\vartheta_1}(t_i)\right);d_i\right),$$

and $\hat{\vartheta}_{1n}$ must fulfill the following

$$\sum_{i=1}^{n} \frac{\partial}{\partial\vartheta_1} Q\left(\varphi\left(\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1),\hat{\phi}_{\vartheta_1}(t_i)\right);d_i\right) = 0. \tag{15}$$

As an example, set $\varphi(u,v) = (u \times v)$, the ACD representation, and $\mu = -\theta^{-1}$ and $V(\mu) = \mu^2$ (the exponential distribution). Then (13) corresponds to the log-likelihood function from an exponential conditional distribution with an ACD representation, i.e.

$$-\sum_{i=1}^{n}\left[\log\left\{\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1)\times\phi(t_i)\right\} + \frac{d_i}{\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1)\times\phi(t_i)}\right] \tag{16}$$

and the first order condition (14), takes the explicit form

$$\hat{\phi}_{\vartheta_1}(\tau) = \frac{\sum_{i=1}^{N} K\left(\frac{\tau-t_i}{h}\right)\frac{d_i}{\psi(\bar{d}_{i-1},\bar{y}_{i-1};\vartheta_1)}}{\sum_{i=1}^{N} K\left(\frac{\tau-t_i}{h}\right)}. \tag{17}$$

Since a closed expression for the parametric part is not available, an iterative algorithm must be used. Now, instead assume that $\varphi(u,v) = \exp(u+v)$, then (13)

corresponds to the log-likelihood function from an exponential distribution with a Log-ACD representation and then and the first order condition (14), takes the explicit form

$$\hat{\phi}_{\vartheta_1}(\tau) = \log\left\{ \frac{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right) \frac{d_i}{\exp\left\{\psi(\bar{d}_{i-1}, \bar{y}_{i-1}; \vartheta_1)\right\}}}{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right)} \right\}. \tag{18}$$

In some situations, it might be also of interest to use a density function that does not belong to the family of exponential functions. This could be the case when we are interested more than in the values of the estimated parameters in density forecasts. In this case, it is possible to perform estimation under these distributions through the use of standard maximum likelihood techniques. Of course it is straightforward to show that under correct specification of the density the results we show further hold.

Therefore for the generalized gamma with parameters $(1, \gamma, \nu)$ by maximizing the corresponding (smoothed) log-likelihood function we obtain the following nonparametric estimator for the seasonal component in the Log-ACD setting

$$\hat{\phi}_{\vartheta_1}(\tau) = \frac{1}{\gamma} \log\left\{ \frac{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right) \left(\frac{d_i}{\exp\left\{\psi(\bar{d}_{i-1}, \bar{y}_{i-1}; \vartheta_1)\right\}}\right)^{\gamma}}{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right) \nu} \right\}. \tag{19}$$

Note that we attain the nonparametric seasonal estimator using the Weibull distribution when $\nu = 1$ and it coincides with the estimator in the Burr case. Finally, the previous expressions have been obtained by assuming an intradaily seasonal component. However, it is also possible to extend it to several seasonal effects. For example, if we consider intraweekly seasonal effects, then we might identify five different seasonal patterns corresponding to each day of the week. In this case, for $s = 1, \cdots, 5$, we have in the exponential and the Log-ACD representation

$$\hat{\phi}_{\vartheta_1}(\tau) = \log\left\{ \frac{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right) \frac{d_i}{\exp\left\{\psi(\bar{d}_{i-1}, \bar{y}_{i-1}; \vartheta_1)\right\}} I\left(t_i \in \Delta_s\right)}{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right) I\left(t_i \in \Delta_s\right)} \right\} \tag{20}$$

where Δ_s is a subset in $[a, b]$ that contains τ. Hence that for any distribution we consider, the non parametric seasonal curve is estimated by nothing else that a transformation of the Nadaraya-Watson estimator of the non parametric regression on the duration adjusted by the long-run component over the curve. The following results are needed to make correct inference for the unknown parameters of the Log-ACD model. In the sequel, we assume that the regularity conditions that are assumed in the Appendix take effect. Then the following results are shown in the Appendix:

Theorem 1 *Under the conditions stated in the Appendix, and if* $h \to 0$ *and* $nh \to \infty$ *then,*

$$\sup_{\theta} \sup_{\tau} |\hat{\phi}_{\vartheta_1}(\tau) - \phi(\tau)| = o_p\left(n^{-1/4}\right)$$

as n tends to infinity.

Theorem 2 *Under the conditions stated in Theorem 1 then,*

$$\sqrt{n}\left(\hat{\vartheta}_{1n} - \vartheta_1\right) \rightarrow_d \mathbf{N}\left(0, \Sigma_{\vartheta_1}^{-1}\right),$$

where

$$\Sigma_{\vartheta_1} = E\left(\frac{\partial^2}{\partial\vartheta_1\partial\vartheta_1^T} Q\left(\varphi\left(\psi(\bar{d}, \bar{y}; \vartheta_1), \phi(t)\right); d\right)\right),$$

as n tends to infinity

5 Application to the trade duration process of a stock in an order driven market

5.1 Data and transformations

In this section we apply the model proposed to a trade duration process. Data are trades during january-march 1998 of Bankinter, a medium size spanish bank traded at Bolsa de Madrid. This stock exchange market is an order driven market and thus it works as some of the most important stock markets in continental Europe like Paris, Brussels or Milan. In a purely order driven market, there is not market maker and all the orders are entered in the order book. When and buying and a selling order match the order is executed. These orders can be either limit orders or market orders.

Our database is a trade database and thus we are not able to examine wether an order comes from a bid or and ask, or a limit or a market order. As we will see later on, the difference between limit and market orders is crucial when explaining why there occur several trades in the same moment, that is, why there are durations equal to zero.

From the original data base two transformations are required. The first has to do with the opening effect while the second one is the already explained way to deal with durations zero.

When a trading day begins, before opening there is an auction in order to fix the opening trading price. Once the price is fixed, all the remaining orders in the auction stay, not being possible to introduce new orders or cancel the existing ones. When the market opens all the orders from the auction are executed in the first minutes. Therefore these trades are not informative about the dynamic of the process and they can be eliminated. Recent studies have eliminated the first half hour of the day for avoiding the effects of the auction in the trading day. Since the moment of time in which the auction orders are traded varies every day, we believe that adopting this approach we lose informative durations. Thus we adopt the "second price" strategy, i.e. consider that the trading day begins from the second price since all the orders traded with the first price correspond to the

orders of the preopening auction. This data transformation has an important effect on the durations equal to zero. Figure 1 reflects this effect. It is the number of durations equal to zero every ten minutes from the opening to the closing including the first trading day price (left plot) and excluding it (right plot). In the case that we include the first price trades it is clear that we will increase artificially the number of trades as well as the number zeros in the sample. Moreover the amount of first price trades is important. In our sample it represents 9.32% of all the trades.

With respect to trades that occur at the same moment of time and are not due to preopening auction, previous studies have assume that they come from a trader that wants to buy or sell a big volume and hence trader splits the order in small blocks that are send it to the order book producing quick trades of some or all of the split orders. Under this assumption, these studies eliminate these trades and thus no durations zero remains in the sample. This trading phenomena can be true in some cases but not in all. Indeed another feasible, and certainly logic, explanation is that these durations zero occur because retail traders post small limit orders being the limit price a round price. In order to verify this conjecture we take a look to figure 2. It represents the number of durations equal to zero (y axe) for all observed prices (x axe). It seems that as a round price happens, for example 1000 pesetas (6.04 euros), the number of trades increases and thus the number of durations zero also increases. This increasing of durations zero does not only happens around the "very round" prices. All the small pikes that can be seen in the figure correspond to prices which are multiples of 50 pesetas (0.3 euros), two times the tick. This confirms the hypothesis that almost all the durations zero occur in round prices and thus they are caused by retail traders that post limit orders at these particular prices. Therefore the durations equal to zero that are not caused by the first trading day price are not deleted but substituted by $1/J$, where J is the total number of successive durations, and the strictly positive duration that follows the durations zero is replaced by its value minus one, as explained in previous section.

5.2 Descriptive analysis

Figures 3 and 4 are the observed durations, the autocorrelogram and a kernel estimate of the density. In figure 3 there is also a piecewise constant curve (the dashed line) indicating the day and the week. The lowest piecewise corresponds to monday and it is increasing up to friday and then it decreases again corresponding to monday of next week. Obviously if one day is holiday there is not piecewise line for that day. In order to see clearly the intradaily seasonal pattern we just show the first month, january 1998. From this figure we can see that everytime that a new day begin there is a decreasing of durations while it increases during the day, indicating the intradaily seasonality. Left plot of figure 4 confirms this feature. It is the autocorrelogram. Even if we should use this plot only for illustrative

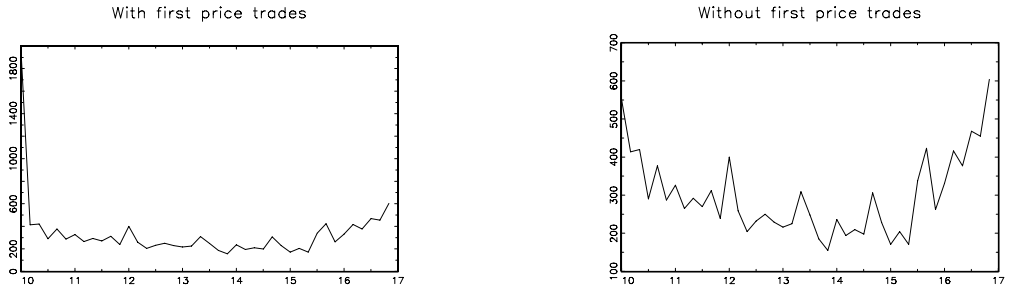


Figure 1: Intradaily seasonality of durations zero. Left including first trading day price. Right excluding it.

purposes,[4] one sees that there is a clear seasonal pattern. Finally right plot of figure 4 shows a kernel estimate of the density. It seems that the density has an asymptote at zero. It implies that, in principle, the exponential distribution should not give good results and the Burr distribution can be redundant in the sense that the second parameter should not be significative and therefore a priori the correct distributions could be either Weibull either generalized gamma.

In table one there is a brief descriptive analysis of some basic measures. Numbers in parenthesis are the same statistics but eliminating durations zero. The basic insights that can be extracted from this table are: durations (with and without zeros) are overdispersed and highly autocorrelated as it was expected given that they are financial processes. The number of durations equal to zero is very significative, 26.5% percent of the total. Eliminating them implies that the dynamical properties of the process change. For example the Q-statistics are higher when only considering strictly positive durations.

Since the aim of the paper is about (intradaily) seasonality, it is worthwhile compute the diurnally component (i.e. the function $\phi(\tau)$ used to seasonally adjust data). Up to now this function has been specified by means of cubic splines

$$\hat{\phi}(\tau) = \sum_{j=1}^{J} \mathbf{1}_{[\Delta_j \leq \tau < \Delta_{j+1}]} \left[a_j + b_j \left(\tau - \Delta_j \right) + c_j (\tau - \Delta_j)^2 + d_j (\tau - \Delta_j)^3 \right],$$

where Δ_j are the knots and $\mathbf{1}_{[\Delta_j \leq \tau < \Delta_{j+1}]}$ is an indicator function for the $j+1th$ segment. We introduce a second estimator which is a standard Nadaraya-Watson

[4] The use of autocorrelograms when dealing with point processes has not an exact meaning. That is, autocorrelation of order 100 implies that the relation between one observation and 100 before is always the same, but one of the main characteristics of tick-by-tick data is that for a given time interval (for example one hour) the number of observations differs and so the seasonality found in the autocorrelogram is just illustrative. A possible alternative to the autocorrelogram is the variogram extensively used in geostatistics and introduced in the analysis of tick-by-tick data by Hillman and Salmon (2001).

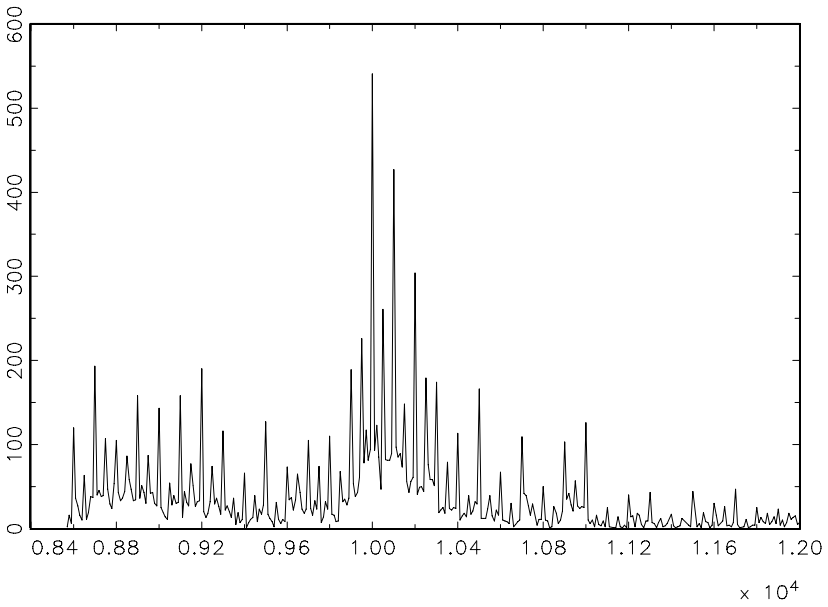


Figure 2: Number of durations zero for every price

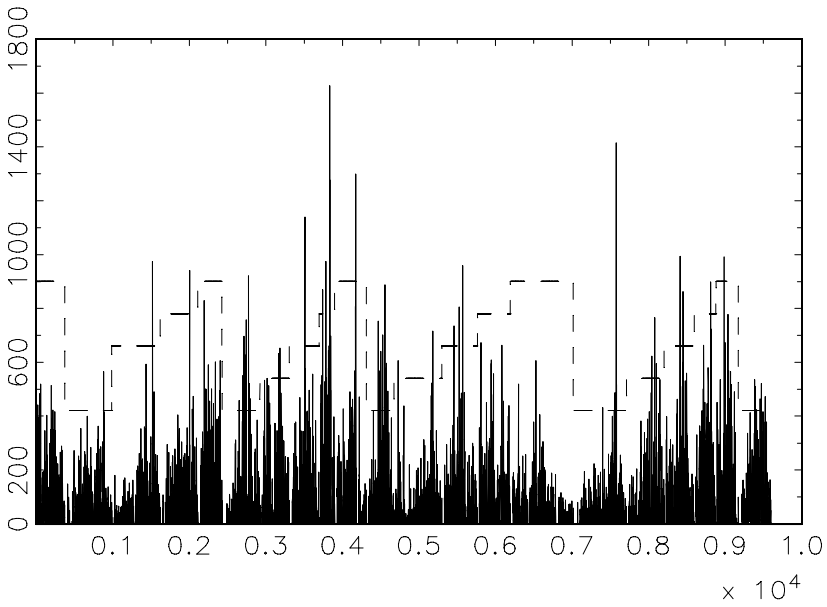
estimator

$$\hat{\phi}(\tau) = \frac{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right) d_i}{\sum_{i=1}^{N} K\left(\frac{\tau - t_i}{h}\right)}.$$

With respect to the later estimator, the time variable is the number of cumulative seconds from midnight every day. The kernel chosen is the quartic and the bandwidth is $2.78\sigma N^{-1/5}$ where σ is the standard deviation of the data and N the number of observations. With respect to the former estimator, the nodes are set every hour and the smoothing parameter is 0.01, as used in previous studies. Figure 5 represents the two diurnally estimators for the mean day and for the five days of the week and adjusting data excluding durations zero.

From this figure it seems that the variation of the daily seasonal pattern is not significative across days. Also the Nadaraya-Watson estimator seems to be smoother than the piecewise cubic spline but the later varies more within a day since it ranges from zero to approximately 90 while the Nadaraya-Watson ranges from approximately 30 to approx 90. [5]

[5] The same exercise has been done including durations zero and results are very similar. This detail is important since as it will be showed later there are remarkable differences between the estimated curves when including and excluding the durations zero



Solid line are observed durations. Dashed line represents the days of the week. Each piecewise is day and it is increasing from monday up to friday. The scheme is repeated every week. Only January has been plotted because representation purposes

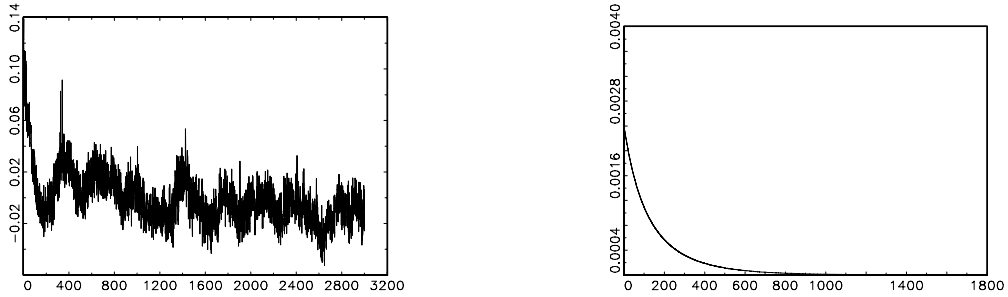
Figure 3: Observed duration and day of the week

5.3 Estimation

We now proceed on estimation. Using the estimation method proposed in the former section, we estimate the parameters of a Log-ACD model and the seasonal component under three alternative specifications for the distribution of the conditional durations: exponential (QMLE), Weibull and generalized gamma. We proceed as well, for comparative purposes, with estimation on raw data, adjusted for seasonality and with and without durations zero.

Results are in tables two and three. The first column is the estimation result when we do not consider seasonality. In the next two columns the intradaily and the intraweekly nonparametric estimators respectively are included. Last column is the result when we previously adjust durations by the Nadaraya-Watson estimator.

The mean equation parameter values, (ω), α and β are the expected according to the properties of a financial duration process. The model is stationary and in all cases models capture the long memory property. Notice that ω is not present in the estimation with seasonal component since the nonparametric curve plays



Density estimated non parametrically with a Gamma Kernel. See Chen (2000). The bandwidth is $(0.9\sigma N^{-0.2})^2$ where σ is the standard deviation of the data and N the number of observations.

Figure 4: Autocorrelogram and Marginal Density

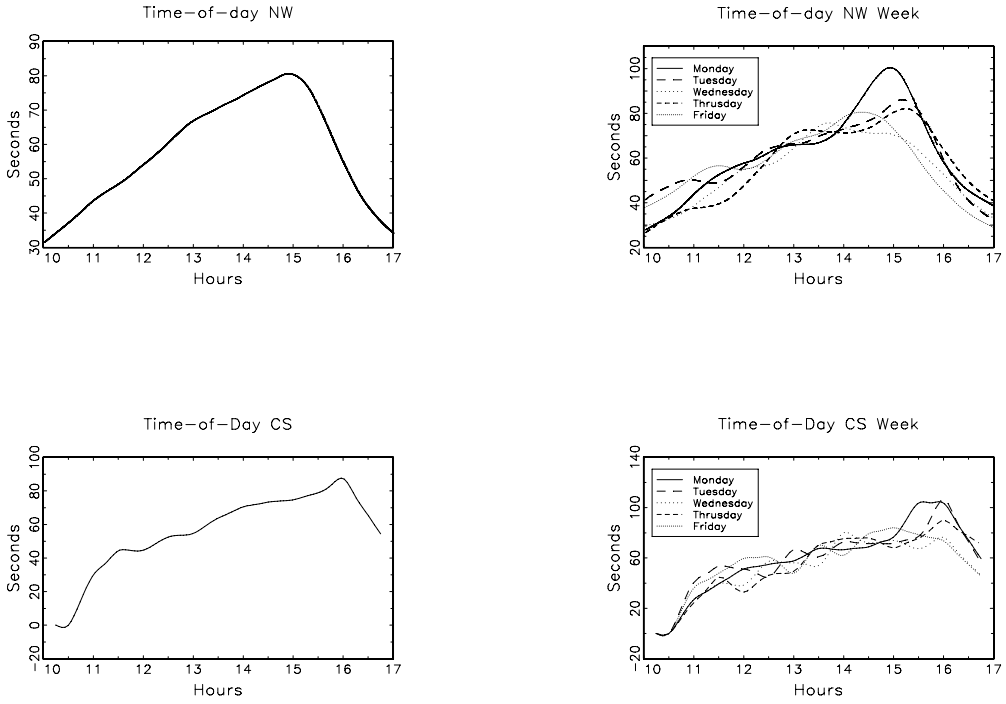
Table 1: Information on Duration Data

No. Days	No. Obs	No. Durations=0	% Durations=0
61 (61)	27298 (20067)	7231 (0)	26.5 (0)
Mean	S.d	Q(1)	Q(10)
55.82 (75.94)	102.28 (112.711)	364.57 (457.58)	2413.3 (2881.8)

Descriptive statistics for the trade durations of Bankinter during January - March 1998. Durations are measured in seconds. $Q(k)$ is the Ljung-Box statistic for autocorrelation of order k.

the role of a time-varying parameter.

With respect to the parameters related with the distribution, when the seasonal component is considered the parameters γ and ν for the generalized gamma increases and decreases respectively. This change can be explained in terms of hazard functions since it is the most important function when dealing with durations. Left plot of figure six shows the hazard functions for the generalized gamma distributions when considering the seasonal component (dashed line) and when ignoring it (solid). Although small, in this plot we can see the effect of considering or not the seasonal term. The hazard function is shifted down (from the solid to the dashed line) when considering the seasonal component. This is due to the following: when getting rid of the seasonal component in the long-run term we are excluding a part of the high activity in the opening and the closing, related with the shorter durations. Equivalently for the lunch time: it is expected that a part of the low trading activity is captured by the seasonal component, related with the longest durations. Therefore the seasonal term will capture a proportion of the lowest and the highest trading activities implying that the hazard function, or



NW stands for Nadaraya-Watson and CS for Cubic Splines

Figure 5: Diurnally component

the instantaneous probability, will decrease and, because the construction of the hazard functions, they also decrease for medium durations. This is why the hazard function shifts down when including the seasonal curve.

Right plot can be used for looking at the differences between distributions. Solid line represents the Weibull hazard function while dashed line is the generalized gamma. We estimate without zeros (the inserted window is a zoom of the area close to the origin). The hazard function of the generalized gamma is above the hazard function of the Weibull. It means that the generalized gamma distribution increases the instantaneous probability of a trade. Finally, remark that as the distribution function becomes more flexible, the changes in the hazard function when estimating with and without seasonal component increases. For example, for the exponential the hazard function is equal through any specification (since it is constant and equal to one), for the Weibull case it varies but very slightly while for the generalized gamma changes are relevant as already explained.

With respect to the seasonal curve, figure 7 shows the intradaily and intraweekly seasonal patterns when using a Weibull distribution for the estimation with and without zeros (bottom and top plots respectively). The first thing that draws the attention is the different shape of the estimated curves by including

Table 2: Estimation Results excluding durations zero

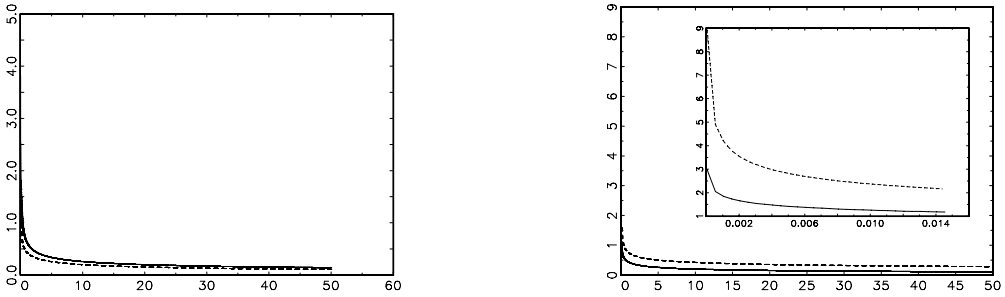
		No Seaso	Intraday	Intraweek	NW
Exp	ω	0.0809 0.0056			0.0196 0.0014
	α	0.0391 0.0016	0.0248 0.0017	0.0235 0.0018	0.0209 0.0015
	β	0.9506 0.0024	0.9703 0.0024	0.9724 0.0024	0.9749 0.0021
Weibull	ω	0.0915 0.0088			0.0234 0.0023
	α	0.0439 0.0025	0.0285 0.0027	0.0268 0.0027	0.0249 0.0025
	β	0.9443 0.0037	0.9655 0.0038	0.9680 0.0039	0.9694 0.0035
	γ	0.7357 0.0047	0.7410 0.0048	0.7421 0.0048	0.7406 0.0048
GG	ω	0.1002 0.0101			0.0261 0.0027
	α	0.0472 0.0024	0.0308 0.0030	0.0288 0.0031	0.0276 0.0029
	β	0.9398 0.0043	0.9622 0.0044	0.9651 0.0044	0.9658 0.0042
	γ	0.5937 0.0213	0.6181 0.0222	0.6252 0.0225	0.6202 0.0222
	ν	1.4313 0.0861	1.3517 0.0804	1.3286 0.0790	1.3429 0.0796

Estimation results ignoring the seasonal behaviour (termed No Seaso), with the non-parametric estimator proposed accounting for the intradaily and the intraweekly pattern (termed Intraday and Intraweek respectively) and using a pre-seasonal adjustment by means of Nadaraya-Watson (NW). Exp, Weibull and GG stand for exponential, Weibull and Generalized Gamma distributions respectively. Numbers in regular character are the estimated parameters and in tiny are heterokedastic-consistent standard deviations.

Table 3: Estimation Results including durations zero

		No Seaso	Intraday	Intraweek	NW
Exp	ω	0.4209 0.0187			0.1149 0.0041
	α	0.0696 0.0023	0.0731 0.0025	0.0751 0.0025	0.0734 0.0025
	β	0.8528 0.0060	0.8204 0.0074	0.8129 0.0075	0.8137 0.0077
Weibull	ω	0.5766 0.0338			0.2069 0.0096
	α	0.1259 0.0053	0.1279 0.0055	0.1292 0.0055	0.1276 0.0055
	β	0.7812 0.0112	0.7353 0.0129	0.7434 0.0133	0.7515 0.0129
	γ	0.5117 0.0028	0.5158 0.0028	0.5168 0.0028	0.5152 0.0028
GG	ω	-0.354 0.3160			-0.489 0.8623
	α	0.2096 0.0066	0.1537 0.0034	0.1496 0.0032	0.2060 0.0066
	β	0.6678 0.0130	0.7023 0.0023	0.7004 0.0023	0.6495 0.0138
	γ	0.0382 0.0002	0.7253 0.0203	0.7309 0.0204	0.0383 0.0002
	ν	146.24 0.1948	1.1235 0.1259	1.1137 0.1262	146.08 0.4274

For explanation see previous table



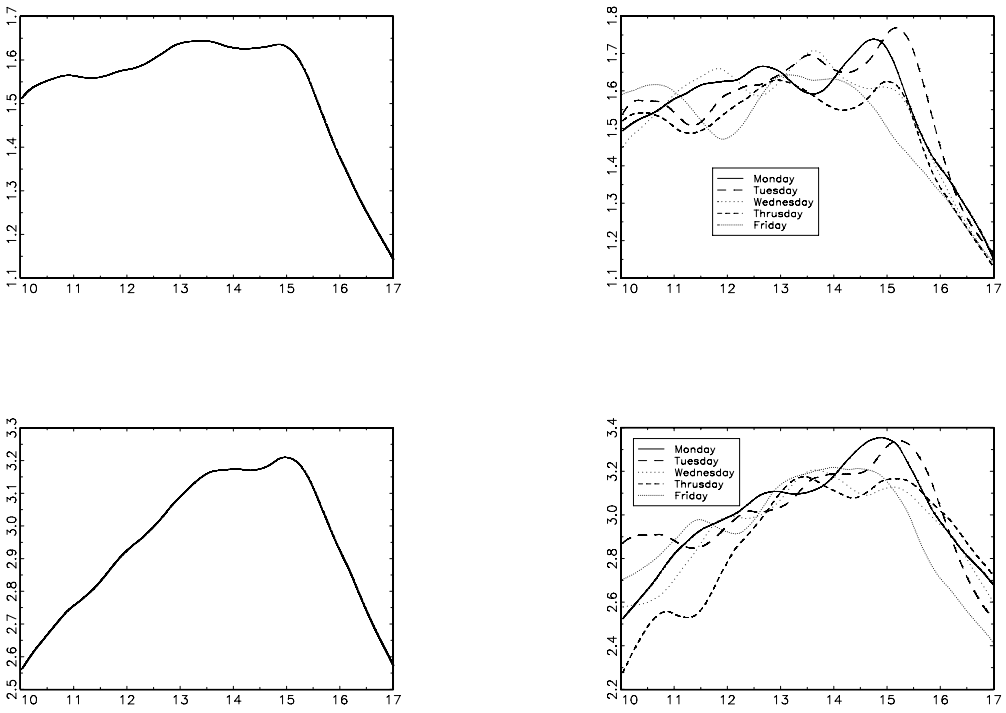
Left plot is the estimated hazard functions for the estimation without durations zero and with and without intradaily seasonal component (dashed and solid lines respectively) for the generalized gamma distribution. Right plot is the estimated hazard functions for the Weibull and the generalized gamma distributions (solid and dashed lines respectively) without durations zero. The inserted window shows a zoom of the graph close to the origin.

Figure 6: Estimated hazard functions

and excluding durations zero. Although they have the same inverted U shape, differences come from the intensity of the seasonality at different periods of the day. It is particularly remarkable at the beginning of the day. In the bottom plots the deterministic seasonality increases sharply at the beginning of the day while it is not the case of the top estimated curve. It means that at the beginning of the day there exists a certain dynamics that is captured by the parametric part in the semiparametric estimation only when excluding durations zero. A comparison can be done whith figure 5. The ad hoc seasonal patterns (including and excluding durations zero) are very similar to the estimated curve when including durations zero. It means that when including durations zero the ones produced in the first half of the day are not informative and hence they are captured by the seasonal curve while it is just the contrary for the durations zero observed at the end of day since the second half day seasonal pattern in similar in any plot. This permit us to conjecture that the exogenous information occurred during the period in which the market is closed is not informative of the stochastic part of the process while the flow of information, either exogenous or endogenous, that arrives to the market when it is open matters.

After 13:00 there are not remarkable differences. Up this time traders go for lunch and just before they take positions, increasing again the trading intensity. Traders lunch and the market remains relatively constant up to a bit before 15:30 when NYSE and NASDAQ preopen and then the market becomes very quickly very active an the trading activity increases and this increase does not stop up the closing at 17:00.

Additionaly, there is not a significative intraweekly pattern since the intradaily seasonality between days of the week are very similar. Finally, although it is not



Top plots are the estimated seasonal intradaily and intraweekly components when excluding durations zero. Bottom plots are the equivalent but including them. Weibull distribution is used

Figure 7: Estimated seasonal curves

showed the seasonal curve is almost identical for any of the three distributions, meaning that it is "robust" to the distribution of the parametric part of the model.

5.4 Diagnosis

For testing the accuracy of the model we use density forecast. This technique is based on the calculation of the probability integral transform and then test wether it is $i.i.d$ and uniformly distributed using histograms and autocorrelograms. It were introduced by Diebold et al. (1998) in the context of GARCH models and it has extensively used by Bauwens et al. (2000) for comparing different financial duration models. This technique is specially useful for evaluating the forecasting performance of different non nested models although it can be used as well for nested models.

Basically it works as follows: $\{f_i(d_i \mid \mathcal{H}_i)\}_{i=1}^m$ a sequence of one-step-ahead density forecasts produced by the model and by $\{p_i(d_i \mid \mathcal{H}_i)\}_{i=1}^m$ the sequence of densities defining the data generating process governing the duration series

d_i. It can be showed that the correct density will be preferred by all forecast users regardless of their loss functions and hence it makes sense to test wether $\{f_i(d_i \mid \mathcal{H}_i)\}_{i=1}^m = \{p_i(d_i \mid \mathcal{H}_i)\}_{i=1}^m$.

This test is done using the probability integral transform

$$z_i = \int_{-\infty}^{d_i} f_i(u)du,$$

that must be *i.i.d.* and uniformly distributed under the correct density. Hence when assuming some mean equation and some distribution both independence and uniformity of the estimated density can be checked.

In order to test uniformity, it can be done easily using a histogram based on an empirical z sequence. If the density is correctly specified the histogram should be statistically flat. For the independence checking, autocorrelation functions of various centered moments can reveal some dependency. For further details see the two above references.

Figures 8 and 9 show the histograms of z and autocorrelograms of $(z_i - \bar{z})$. Figure 8 are, from top to bottom, the density forecast results when estimating without durations zero and the three distributions. Last row when considering durations zero and the generalized gamma distribution. All estimations with the intradaily component and looking at the out-of-sample performance. On the contrary figure 9 are the estimations with the generalized gamma, the Nadaraya-Watson estimator and with and without durations zero. We do not show the autocorrelograms for other centered moments and using the intraweekly component since results are similar in all cases.

From these figures some comments arise: First in general the mean equation captures correctly the dynamic in any case since most of the autocorrelations remain in the 90% cofindence bands. This result is also found in Bauwens et al. (2000) where they shown that the mean equation choice in not crucial for determining the accuracy of the model. There is some residual autocorrelation when durations zero are included and when the seasonal curve is not estimated jointly. Secondly there is in general a huge difference between the estimation with and without the durations zero. This is caused probably by the way in which durations zero are dealt. As explained in previous section we did not expect good results and thus we let the improvement on the treatment of these data for future research. Nevertheless it is worthwhile explain why this shape. Indeed a similar shape (to the one of the last row of figure 8) has been found in Bauwens et al. (2000) when dealing with price durations and previously adjusting data by means of a cubic spline. The considered distributions are not able to account for durations smaller than one which is probably due to its high percentage with respect to the whole sample. This is represented in the histogram with a very small frequency for $0 < z < 0.05$ and hence this lack of values at this range provokes an over representation on the following bins.

With respect to the distributional assumption, as expected the exponential distribution does not make a good job while there other two behaves much better,

specially the generalized gamma. Even when the durations are preadjusted, the histogram, although significantly worse than in the joint estimation, does not look like very bad but it is far from uniformity since large and small durations are under represented.

Related with the inclusion of the seasonal component differences are clear. When it is included in the estimation, forecasting results are much better and z is uniformly distributed (in the case of no durations zero) and hence we assert that only when including in the estimation the seasonal component the forecasted probability integral transform is $i.i.d.$ and uniformly distributed.

6 Conclusions

When dealing with tick-by-tick data, there exists seasonal patterns. One of the most significative is the intradaily seasonality, that is, the deterministic pattern that all the tick-by-tick variables shown through the day. Since data are irregularly spaced, that is they are a point process, the analysis of the seasonality cannot be done using standard tools. On the other hand if the point process can be decomposed in components, such as long-run (accounting for the dynamics)and short-run (for the seasonality) , estimation of the components in steps is not eficcient since they are not orthogonal. In this paper we have proposed a semiparametric modeling strategy. The dynamics of the process are modeled parametrically while the seasonality is a nonparametric curve. This semiparametric approach is justified since we are interested in the analysis of the dynamics of the process but accounting for the existence of a very strong seasonal component.

Simultaneous estimation of the parameters of the dynamic component and the seasonal curve is performed by a modification of the method suggested in Severini and staniswalis (1994). An explicit from the estimator of the seasonal curve is obtained and it depends on the underlying distribution that is assumed. Moreover we show that the nonparametric estimator is consistent and that the estimated parameters, given the estimated curve, are asymptotically consistent, efficient and normally distributed.

Some discussion is introduced about how to deal with durations zero since they are often observed. Finally we apply the above proposed methodology to the trade process of a Spanish bank traded in Bolsa de Madrid. This stock market is a purely order book market as many of the continental Europe stock exchanges. Results show that the seasonality is very strong having different behaviour in the opening and the closing times as well in east-coast USA stock markets' opening.

Some extensions are possible. The most important is a better way to deal with durations zero. Using the approach proposed here does give nice results. Going further in dynamical hurdle models or censoring is future research. On the other hand this analysis can be useless since due to technologies improvements the time measure will be smaller and hence there will be a moment in which there will not be durations zero. Actually some stock exchange markets sell tick-by-tick data in

centesimal seconds.

Some other extension is the analysis of any other tick-by-tick variable using this semiparametric approach. For example volatility. In general once the dynamic component is specified as well as the distribution, the semiparametric approach proposed here can be used giving as result consistent and efficient parameters.

Appendix

Definitions and assumptions

In order to prove the results claimed in Theorems 1 and 2 we need to establish some definitions and assumptions. The proofs follow the same lines as in Severini and Staniwallis (1994).

(A.1) The random variable t takes values in a compact set $\mathbf{T} \subset R$. The marks y take values in a compact set $\mathbf{Y} \subset R^p$.

(A.2) The observations $\{(d_i, y_i, t_i)\}_{i=1,\cdots}$ are a sequence of stationary and ergodic random vectors.

(A.3) ϑ_{10} takes the values in the interior of Θ, a compact subset in R^p and ϕ takes the values in the interior of Λ, a compact subset of R.

$$\Lambda = \left\{ f \in C^2[a,b] : f(t) \in \operatorname{int}(\Lambda) \quad \text{for } \forall t \in [a,b] \right\}.$$

(A.4) Let Ξ be a compact subset of R such that $\varphi\left(\psi(\bar{d}, \bar{y}; \vartheta_1), \phi(t)\right) \in \Xi$ for all $t \in \mathbf{T}$, $y \in \mathrm{T}$, $\vartheta_1 \in \Theta$ and $\phi \in \Lambda$.

(A.5) The matrix

$$\Sigma_{\vartheta_1} = E\left(\frac{\partial^2}{\partial \vartheta_1 \partial \vartheta_1^T} Q\left(\varphi\left(\psi(\bar{d}, \bar{y}; \vartheta_1), \phi(t)\right); d \right) \right)$$

is positive definite.

(B.1) The kernel function $K(\cdot)$ is of order $k > 3/2$ with support $[-1, 1]$ and it has bounded $k+2$ derivatives.

(B.2) For $r = 1, \cdots, 10 + k$ the functions $\partial^r \varphi(m)/\partial m^r$ and $\partial^r V(\mu)/\partial \mu^r$ exist and they are bounded in their respective supports.

(B.3) d is a strong mixing process where the mixing coefficients must satisfy for some $p > 2$ and r being a positive integer

$$\sum_{i=1}^{\infty} i^{r-1} \alpha(i)^{1-2/p} < \infty.$$

Furthermore, for some even integer q satisfying $\frac{(k+2)(3+2k)}{(2k-3)} \leq q \leq 2r$

$$E\left|d\right|^{q} < \nu,$$

where ν is a constant not depending on t.

(B.4) The conditional density of t, given the information set I_{i-1}, $f(t)$, and the conditional density of d given t and I_{i-1} has $k+2$ bounded derivatives uniformly in $t \in \mathbf{T}$, $y \in \mathbf{Y}$ and $d \in \mathbf{D}$.

(B.5) Let

$$M\left(\eta;\vartheta_1,t\right) = E\left\{\frac{\partial}{\partial\eta}Q\left(\varphi\left(\psi(\bar{d},\bar{y};\vartheta_1),\eta\right);d\right)\left|\bar{y},\bar{d}\right.\right\}.$$

For each fixed ϑ_1 and t, let $\phi_{\vartheta_1}(t)$ the unique solution to $M\left(\eta;\vartheta_1,t\right)=0$. Then for any $\epsilon>0$ there exists a $\delta>0$ such that

$$\sup_{\vartheta_1\in\Theta}\sup_{t\in\mathbf{T}}\left|\phi_{\vartheta_1}(t)-\phi(t)\right| < \epsilon$$

whenever

$$\sup_{\vartheta_1\in\Theta}\sup_{t\in\mathbf{T}}\left|M\left(\phi(t);\vartheta_1,t\right)\right| < \delta.$$

(B.6) The sequence of bandwidths must satisfy $h = O(n^{-\alpha})$ where

$$\frac{1}{4k} < \alpha < \frac{1}{4}\frac{q-(2+p)}{q+(2+p)}.$$

Proof of Theorem 1

The proof of this theorem follows the same steps as in the proof of Lemma 5 from Severini and Wong (1992), p. 1784. The bias term must be treated in the same way as they do. With respect to the variance term an additional result must be included to account for the dependence. Consider the following expression

$$\frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{\tau-t_i}{h}\right)\varphi\left(\psi(\bar{d},\bar{y};\vartheta_1),\eta\right) - E\left\{K\left(\frac{\tau-t}{h}\right)\varphi\left(\psi(\bar{d},\bar{y};\vartheta_1),\eta\right)\right\}\right]$$

and define

$$W_i = \frac{1}{h}K\left(\frac{\tau-t_i}{h}\right)\varphi\left(\psi(\bar{d},\bar{y};\vartheta_1),\eta\right) - E\left\{K\left(\frac{\tau-t}{h}\right)\varphi\left(\psi(\bar{d},\bar{y};\vartheta_1),\eta\right)\right\}$$

Then, under assumptions (A.2) and (B.3) the process $W_1,\cdots,W_n$ is strong mixing and therefore theorem 1 from Cox and Kim (1995) applies and the following sequence of inequalities hold. For $\epsilon>0$

$$P\left\{\left|\frac{1}{n}\sum W_j\right| > \epsilon\right\} \leq \frac{E\left[\left(\sum W_i\right)^q\right]}{n^q\epsilon^q} \leq$$

$$\frac{1}{n^q \epsilon^q} C \left\{ n^{q/2} \sum_{i=P}^{\infty} i^{q/2-1} \alpha(i)^{1-2/p} + \sum_{j=1}^{q/2} n^j P^{q-j} \nu^j \right\}$$

for any integers n and P with $0 < P < n$. Then using assumptions (B.1) to (B.6) and proceeding as Severini and Wong (1992) in the proof of Lemma 8, the proof is closed.

Proof of Theorem 2

The proof of this theorem relies consists in verifying conditions I (Identification), S (Smoothness) and NP (Nuissance Parameter) fromm Severini and Wong (1992). Condition NP(a) is the result already shown in Theorem 1. Condition NP(b) (least favorable curve) is inmediate from Lemma 6 of Severini and Wong (1992). This is due to the fact that we assume that the conditional density function belongs to the exponential family. By assuming (A.1) to (A.4) the smoothness condition holds. Finally, assumption (A.5) implies I. Then, using both a Uniform Weak Law of Large Numbers and a Central limit theorem for a stationary and ergodic process (see for example Wooldridge, 1994) propositions 1 and 2 from Severini and Wong (1992) apply and the proof is done.

References

[1] Almeida, A., Goodhart, C. and Payne, R. (1996). "The effects of macroeconomics 'news' on high frecuency exchange rate behaviour," Mimeo LSE/Financial Markets Group.

[2] Andersen, T. and Bollerslev, T. (1997). "Heterogenous Information Arrivals and Return Volatility Dynamics: Unrecovering the Long-Run in High Frequency Returns," *The Journal of Finance*, Vol. LII, n. 3, 975-1005.

[3] Andersen, T. and Bollerslev, T. (1998). "Deustche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies," *The Journal of Finance*, Vol. LIII, n. 1, 219-265.

[4] Baillie, R. and Bollerslev, T. (1990). "Intra-Day and Inter-Market Volatility in Foreign Exchange Rates," *Review of Economic Studies*, 58, 565-585.

[5] Bauwens, L. and Giot, P. (1999). "The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks," *Annales d'Economie et de Statistique*, Vol. 60, 117-149

[6] Bauwens, L. Giot, P. Grammig, J. and Veredas, D. (2000). "A comparison of financial duration models via density forecast," CORE DP 2000/60. Université catholique de Louvain.

[7] Bauwens, L. and Veredas, D. (1999). "The Stochastic Conditional Duration Model: A latent factor model for the analysis of financial durations," CORE DP 9958. Université catholique de Louvain.

[8] Beltratti, A. and Morana, C. (1999). "Computing value at risk with high frequency data," *Journal of Empirical Finance*, 6, 431-455

[9] Bollerslev, T. and Domowitz, I. (1993). "Trading Patterns and Prices in the Interbank Foreign Exchange Market," *The Journal of Finance*, Vol. XLVIII, 4, 1421-1443.

[10] Camacho, C. and Veredas, D. (2001). "Random Aggregation in ACD models when the stopping time is either endogenous or exogenous," Mimeo CORE. Université catholique de Louvain.

[11] Chen, S.X. (2000). "A Beta Kernel Estimation for Density Functions," *Computational Statistics and Data Analysis*, 31, 131-145.

[12] Cox, D.R. and Kim, T. Y. (1995). "Moment bounds for mixing random variables useful in nonparametric function estimation," *Stochastic processes and their applications*, 56, 151-158.

[13] Diebold, F.X., Gunther, T.A., and Tay, A.S. (1998). "Evaluating density forecasts, with applications to financial risk management," *International Economic Review* 39, 863-883.

[14] Drost,F.C. and Werker, B.J.M. (2001). "Efficient estimation in semiparametric time series: the ACD model," CentER Discussion Paper 2001-11, Tilburg University.

[15] Engle, R.F., Ito, T. and Lin, W.L. (1990). "Meteor Showers or Heat Waves? Heteroskedastic Intra-Daily Volatility in the Foreign Exchange Market," *Econometrica*, Vol. 58, n. 3, 525-42.

[16] Engle, R.F. and Russell, J.R. (1997). "Forecasting the Frequency of Changes in Quoted Foreign Exchange Prices with the ACD model," *Journal of Empirical Finance* n. 4, 187-212.

[17] Engle, R.F. and Russell, J.R. (1998). "Autoregressive conditional duration: a new approach for irregularly spaced transaction data," *Econometrica* Vol. 66, n. 5, 1127-1162.

[18] Engle, R.F. (2000). "The Econometrics of Ultra High Frequency Data," *Econometrica* Vol. 68, n. 1, 1-22.

[19] Fan, J., Heckman, N. E. and Wand, M. P. (1995). "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions". *Journal of the American Statistical Association*, 90, 141-150.

[20] Gerhard, F. and Haustch, N. (1999). "Volatility Estimation on the Basis of Price Intensities," Mimeo Center of Finance and Econometrics. University of Konstanz.

[21] Ghysels, E., Gouriéroux, C. and Jasiak, J. (1997). "Stochastic Volatility duration models," Working paper 9746. CREST Paris.

[22] Ghysels, E. (2000). "Some Econometric Recipes for High Frequency Data Cooking," *Journal of Business and Economic Statistics* Vol. 8, n. 2, 154-163.

[23] Gouriéroux, C., Monfort, A. and Trognon, A. (1984). "Pseudo Maximum Likelihood Methods: Theory," *Econometrica* Vol. 52, n. 3, 681-700.

[24] Gouriéroux, C., Jasiak, J. and Le Fol, G. (1999). "Intra-Day Market Activity," *Journal of Financial Markets*, 2, 193-226.

[25] Grammig, J. and Maurer K.O. (1999). "Non-monotonic hazard functions and the autoregressive conditional duration model," *The Econometrics Journal*, 3, 16-38.

[26] Harris, L. (1986). "A transaction data study of weekly and intradaily patterns in stock returns," *Journal of Financial Economics*, 16, 99-117.

[27] McCullagh, P. and Nelder, J.A. (1983). *Generalized linear models*, London :Chapman and Hall.

[28] Payne, R. (1996). "Announcement Effects and Seasonality in the Intra-day Foreign Exchange Market," Mimeo, LSE/Financial Markets Group.

[29] Severini, T.A. and Staniswalis, J.G. (1994). "Quasi-likelihood estimation in semiparametric models," *Journal of the American Statistical Association*, 89, 501-511.

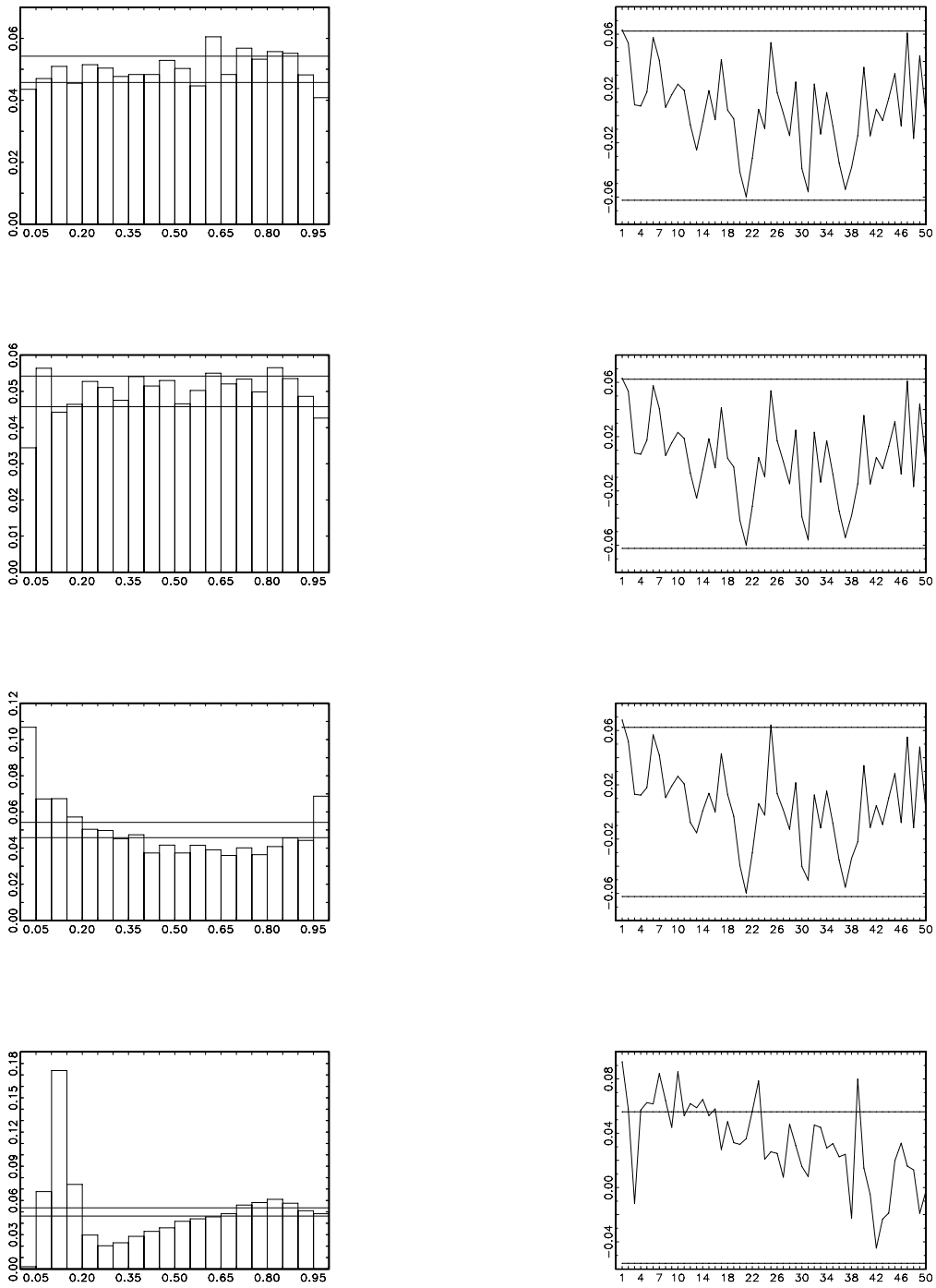
[30] Severini, T.A. and Wong, W. H. (1992). "Profile likelihood and conditionally parametric models," *Annals of Statistics*, 20, 1768-1802.

[31] Staniswalis, J.G. (1989). "On the kernel estimate of a regression function in likelihood based models," *Journal of the American Statistical Association*, 84, 276-283.

[32] Wei, S.X. (1997). "A Bayesian Approach to Dynamic Tobit Models," CORE DP 9781. Université catholique de Louvain.

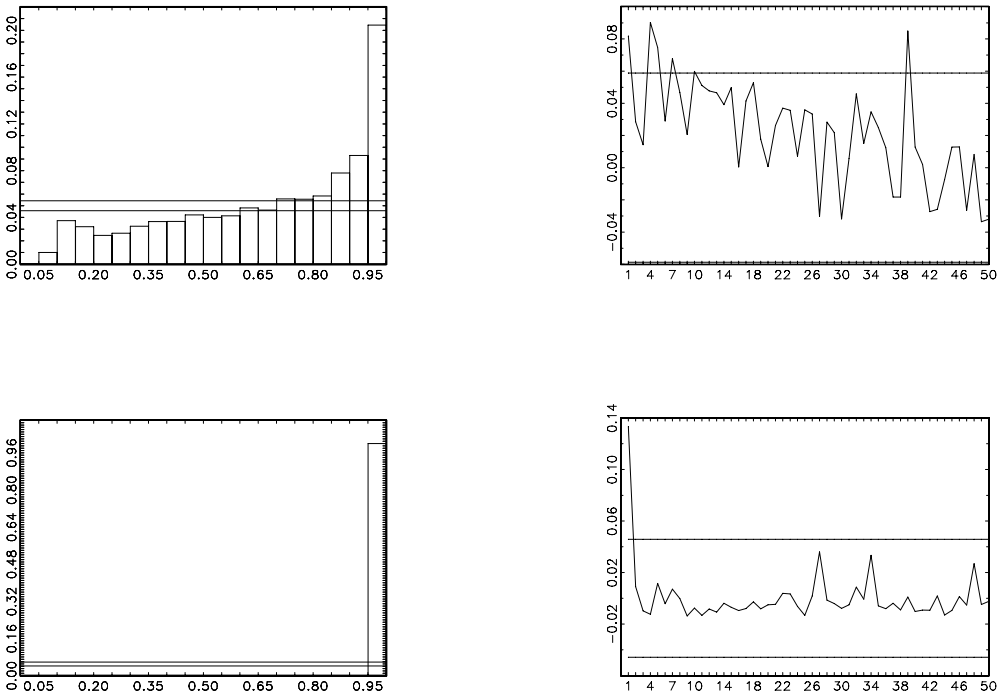
[33] Wooldridge, J.M. (1994). "Estimation and inference for dependent processes," *Handbook of Econometrics*, vol. 4. Engle, R.F. and D.L. McFadden eds. Elsevier Science. New York.

[34] Zhang, M.Y., Russell, J.R. and Tsay, R.T. (1999). "A nonlinear autoregressive conditional duration model with applications to financial transaction data," Mimeo. Graduate School of Business. University of Chicago.



Histograms and autocorrelograms for z. Intradaily component used. Top three without durations zero. Bottom one with durations zero. Specifications from up to down: generalized gamma, Weibull, exponential and generalized gamma.

Figure 8: Density forecast evaluation for raw durations



Histograms and autocorrelograms for z. Adjusted for seasonality using the Nadaraya-Watson estimator and the generalized gamma distribution. Top without durations zero. Bottom with.

Figure 9: Density forecast evaluation for seasonally adjusted durations