BIN models and activities on financial markets

Mamoudou Hassane
Faculty of Economics and Law (FSEJ), University Abdou Moumouni, Niamey BP 12442 Niamey – Niger.
Author's E-mail: hassanemamoudou@hotmail.com

Abbreviations
ACD: Autoregressive Conditional Durations; ARCH: Autoregressive Conditional Heteroscedasticity; ARMA: Autoregressive Moving Average; DGP: Data Generating Process; GARCH: General ARCH; MLE: Maximum Likelihood Estimation

INTRODUCTION

It is usual to find time series consisting of count data. Such series record the number of events of a particular type occurring in a given interval. Since the data are considered to be non-negative integers, a model based on a normal distribution is not appropriate, although it might provide a reasonable approximation if the number of events observed in each time period is relatively large. Then for small numbers, the good distribution is a binomial process, but for a large number of observations the appropriate distribution is the Poisson. Therefore, the Poisson process should be used in the case of count data. This last case allows the formulation of the BIN models.

The recent rise in the availability of high-frequency financial data has seen an increase in the number of studies focusing on areas of classification and modeling of financial markets at the ultra-high frequency level. The development of models that are able to reflect the various observed phenomena of real data is an important step towards obtaining a full understanding of the fundamental stochastic processes driving the market.

The statistical properties of high-frequency financial data and market micro-structural properties were studied by means of different tools including phenomenological models of price dynamics and agent-based market simulations (Goodhart and O’Hara, 1997; O’Hara, 1999; Madhavan, 2000; Scalas et al., 2000; Mainardi et al., 2000; Dacorogna et al., 2001; Raberto et al., 2001; Cincotti et al., 2003; Luckock, 2003; Scalas et al., 2004; Pastore et al., 2010; Ponta et al., 2011, Ponta et al., 2011; Mandelbrot, 1963; 1997; M’uller et al., 1990; Mantegna and Stanley, 1995; Gopikrishnan et al., 2000; Hautsch, 2012). Various studies on high-frequency econometrics appeared in the literature using the autoregressive conditional duration models (Engle and Russell, 1997;
Alternative stochastic models have also been proposed, e.g., diffusive models, ARCH-GARCH models, stochastic volatility models, models based on fractional processes, models based on subordinate processes (Cont and Bouchaud, 2000; Chowdhury and Stauffer, 1999; Hardle and Kirman 1995; Levy and Solomon, 1995; Lux and Marchesi, 1999; Stauffer and Sornette, 1999; Yousefsmir and Huberman, 1997). An important variable is the order imbalance. Many existing studies analyze order imbalances around specific events or over short periods of time. For example in Blume et al., (1989), order imbalances are analyzed around the October 1987 crash. Chan and Fong (2000) analyze how order imbalances change the contemporaneous relation between stock volatility and volume using data of about six months. A large body of research examines the effect of the bid-push spread and the impact of the order imbalance on the short-run behavior of prices (Stoll and Whaley (1990); Hauser and Lauterbach (2003); Chordia et al. (2002), Ponzi et al. (2009), Sverencik and Slania (2007), Wyart et al. (2008), Moro et al. (2009), Perell’o et al. (2008), Preis et al. (2011), Kumasresan and Krejc (2010), Zaccaria et al. (2010), Lim and Coggins (2005), Weber and Rosenow (2005), Bouchaud (2005)). Other studies have examined trading activity as measured by the average number of trades in a unit time [Bonanno et al., (2000), Pierou et al. (2000)]. However, aggregation into uniform time intervals may affect the analysis, since choosing a short unit time interval may result in many points with none or very few trades, artificially altering the heteroskedasticity of the process while using a long unit time interval averages out multiple transactions, and the fine timing structure of the data can be lost (Engle and Russell, 1998). For this reason, another important empirical variable is the waiting time between two consecutive transactions (Scalas et al. (2004), Scalas (2006a), Scalas (2006b)). Empirically, in the market, during a trading day the activity is not constant (see Engle and Russell (1997), Engle and Russell (1998) leading to fractal-time behavior (Hudson and Mandelbrot, 2010), Vrobel 2011). Indeed, as a consequence of the double auction mechanism, waiting times between two subsequent trades are themselves stochastic variables (Scalas, 2006a; 2006b; 2007; Politi and Scalas, 2008). They may also be correlated to returns (Raberto et al., 2002) as well as to traded volumes. In the last few years, in order to investigate tick-by-tick financial time series, the continuous-time random walk (CTRW) has been used (Scalas et al., 2000; Masoliver, Montero et al., 2003; Ivanov et al., 2004; Scalas, 2006a; b). It turned out that interorder and intertrade waiting-times are not exponentially distributed.

Therefore, the jump process of tick-by-tick prices is non-Markovian (Scalas et al. (2000), Scalas (2006a), Scalas (2006b)). Bianco and Grigolini applied a new method to verify whether the intertrade waiting time process is a genuine renewal process (Goldstein, Morris and (2004), Embrechts et al. (1997), Press, Flannery and Teukolsky (1992). This was assumed by the CTRW hypothesis by Scalas et al. (2000). They found that intertrade waiting-times do follow a renewal process. Indeed, trading via the order book is asynchronous and a transaction occurs only if a trader issues a market order. For liquid stocks, waiting times can vary in a range between fractions of a second to a few minutes, depending on the specific stock and on the market considered. In Raberto et al. (2002), a study on General Electric stocks traded in October 1999 is described. Waiting times between consecutive prices exhibit 1-day periodicity, typical of variable intraday market activity. Moreover, the survival probability (the complementary cumulative distribution function) of waiting times is not exponentially distributed (Mainardi et al., 2004; Scalas, 2006a), but is well fitted by a Weibull function (Engle and Russell, 1994; 1997; 1998), Raberto et al., 2002, Ponta et al., 2012).

Here, inspired by Bertram (2005), we propose a model based on nonhomogeneous Poisson processes. Bauwens and Veredas (2004) proposed the stochastic conditional duration process which is based on the assumption that the durations are generated by a latent stochastic factor that follows a first order autoregressive process. The model is applied to stock market price-durations, looking at the relation between price durations, volume, spread and trading intensity. Ghysels et al. (2004) introduce the stochastic volatility duration model (SVD) to handle higher-order dynamics in the duration process. They present the distributional properties of the model, and using trade durations of Alcatel, they compared its performance to the performance of ACD models. Gouriéroux and Jasiak (2005) introduced the Autoregressive gamma process (ARG process) which can accommodate complex nonlinear dynamics, and represent processes with long memory more parsimoniously than ACD and SVD processes.

Several authors such as Engle and Russel (1998), Bauwens and Giot (1999) have previously worked on high frequency financial data. These works dealt with the time between the financials events as trades through autoregressive conditional duration (ACD) models, while the so-called BIN models used for count data deal with the number of events of the high frequency data (as trades) during fixed durations.

The autoregressive conditional duration (ACD) model of Engle and Russell (1998) is one of the most important models of the durations in econometric literature. Here $\psi_\tau = E(x_t | F_{t-1})$ is the conditional expected waiting time. In practice Engle and Russell (1998) have used an exponential or Weibull distribution on the $\{\varepsilon_t\}$. Straightforward alternative structures would be to parameterize the $\log \psi_\tau$ instead of the $\psi_\tau$. This formulation called logACD proposed by Bauwens and Giot (1999), allows to avoid making constraints on parameters.

The two types of models are applied to high frequency data, particularly financial data. The ACD models study the
distribution of duration between the events (quote trades, volume or price duration) while the BIN models focus on the distribution of the number of events during a fixed length time. Therefore the two types of models study the two faces of the same reality, but could be considered more complementary than substitue.

The aim of this survey is to study the degree of relevance of BIN(1,1), the autoregressive form of the BIN models, in other words what is the degree of explanation of financial market events. The aim of this paper is therefore to exhibit theoretical and empirical works on the BIN models to prove that these models could achieve similar results as the GARCH models to describe and analyze activities in financial markets. The current models used by financial institutions and financial markets analysts are the ARCH-GARCH models and their derivatives.

**Count data model: BIN models**

We consider processes that produce discrete, isolated events in some interval, possibly multidimensional. We will make inferences about the event rates per unit interval.

Examples: Arrival time series: \( D = \{ t_i \} \) rate \( r(t) = \text{events s}^{-1} \)

If our measurements are trivial, “BIN” events and can only report the number of events in one or more finite intervals. Then the appropriate model is the Poisson counting process.

If our measurements have sufficient resolution for us to measure every individual event, the appropriate model is the Poisson point process.

If the event characteristics are measured with error, it is a point process with error.

If the event rate is constant over the entire interval of interest, the process is homogeneous; otherwise it is inhomogeneous.

Considering an independent Poisson random variable, if \( n_1, ..., n_q \) are independent with \( n_j \sim Po(\lambda_j) \), then the total of all the counts is \( n_1 + n_2 + ... + n_q \sim Po(\lambda_1 + ... + \lambda_q) \) and the given total of the counts are;

\[
(n_1, ..., n_q) \sim Mult(N, p_1, ..., p_q)
\]

where

\[
N = n_1 + ... + n_q \quad \text{and} \quad p_i = \lambda_i / (\lambda_1 + ... + \lambda_q)
\]

\( i = 1, ..., q \).

The conditional distribution is important for the analysis of log-linear models and it leads us to an analysis based on multinomial distribution.

The autoregressive conditional duration (ACD) model of Engle and Russell (1998) is one of the most important models of duration in econometric literature. It was formulated as follows:

\[
x_t = \varepsilon_t \psi_t, \quad \varepsilon_t > 0, \quad E(\varepsilon_t) = 1
\]

where \( x_t = \tau_t - \tau_{t-1}, \quad t = 1, 2, ... \) is the length of time between financial events (trades), and the \( \varepsilon_t \)'s are independent identically distributed (i.i.d.) with \( \psi_t = \alpha + \sum_{j=1}^q \gamma_j x_{t-j} + \sum_{j=1}^q \beta_j \psi_{t-j} \).

Here \( \psi_t = E(\varepsilon_t | F_{t-1}) \) is the conditional expected waiting time. In practice Engle and Russell (1998) used an exponential or Weibull distribution on the \( \{ \varepsilon_t \} \). A straightforward alternative structures would be to parameterize the log \( \psi_t \) instead of the \( \psi_t \). This formulation called logACD proposed by Bauwens and Giot allows the avoidance of making constraints on parameters.

Rydberg and Shephard (2000) showed how to model an asset price \( p(r) \) at time \( r \) using a compound Poisson process thus:

\[
p(r) = p(o) + \sum_{r=1}^{N(r)} z_r,
\]

(1)

where \( \{ N(r) \}_{r \geq 0} \) is a number of trades recorded up until \( r \) and \( z_r \) is the price movement or change associated with the \( t \)-th trade. Rydberg and Shephard (2000) specified \( N(r) \) to be a counting process, modelled as Cox process — that is a Poisson process with a random intensity.

From an economic viewpoint these authors are typically interested in comparing the rate of return on holding the asset with that obtainable by other risky investments (opportunity cost) or riskless interest rate bearing accounts. In order to do this, one has to compute the return over a fixed length of time \( \Delta \geq 0 \). Then these returns will be based around the difference thus:

\[
p_i = p((i + 1)\Delta) - p(i\Delta)
\]

\[
= \sum_{r=1}^{N((i + 1)\Delta)} z_r - \sum_{t=1}^{N(i\Delta)} z_t
\]

\[
= \sum_{t=N(i\Delta)+1}^{N((i+1)\Delta)} z_t.
\]

This shows that the number of trades in the interval \([i\Delta, (i+1)\Delta] \) plays a crucial role. To reflect this, Rydberg and Shephard specified an expression.

---

1 The counting process, which is used in this context, states, that if \( \{ N(r) \}_{r \geq 0} \) is a process with state space \( \mathbb{Z} [(-\infty, \Delta)] \) and non-decreasing right continuous paths, then \( \{ N(r) \}_{r \geq 0} \) is a counting process.
\[ N_i = N(i+1)\Delta - N(i\Delta), \quad (2) \]

which is the number of trades in that time interval\(^2\). This operation called “binning operation” consists of partitioning time into sections and we count the number of trades in that interval. Note that if \( N_i = 0 \), then \( p_i = 0 \), while \( N_i > 0 \) the prices can change. Therefore, \( N_i \) is very important in determining the activities in price level changes. For small values of \( \Delta \), there will be a negligible loss in information in doing this compared to studying the complete record of the \{ \( N(r) \) \} process.

Let the \{ \( N_i \) \} and \{ \( z_i \) \} processes be stochastically independent and covariance stationary assuming that \{ \( z_i \) \} is independent and identically distributed. Then, writing \( F_{i\Delta} \) as the information about the \{ \( N_i \) \} sequence available infinitesimally before time \( i\Delta \) by assuming the moments exist will be:

\[
\begin{align*}
\text{Var}(p_i | F_{i\Delta}) &= E\{\text{Var}(p_i | N_i) F_{i\Delta}\} + \text{Var}\{E(p_i | N_i) F_{i\Delta}\} \\
&= \text{Var}(z_i) E(N_i | F_{i\Delta}) + E(z_i)^2 \text{Var}(N_i | F_{i\Delta}).
\end{align*}
\]

Thus, predicting the variance of the price over the next period of length \( \Delta \) requires modeling the mean and variance of the number of future trades. In practice \( E(z_i) \) will be too small and so what matters in the above setup is really only \( E(N_i | F_{i\Delta}) \).

By setting \( E(z_i) = 0 \) then

\[
\begin{align*}
\text{Cov}(p_i^2, p_{i+1}^2) &= E\{\text{Cov}(p_i^2, p_{i+1}^2 | N_i, N_{i+1})\} + \text{Cov}\{\text{Var}(p_i | N_i), \text{Var}(p_{i+1} | N_{i+1})\} \\
&= \text{Var}(z_i)^2 \text{Cov}(N_i, N_{i+1}).
\end{align*}
\]

Hence volatility clustering can be obtained with the autocorrelation of square price changes being proportional to the counts.

A more specific result is obtained by assuming that \{ \( z_i \) \} has a first order moving average representation (Rydberg and Shephard, 2000). Basically empirical modeling would require the assumptions that \{ \( N_i \) \} and \{ \( z_i \) \} are stochastically independent.

\(^2\) Other derived financial activities could be used as object of count, as quote, and volume to obtain the count data.

**Poisson process**

Under some assumptions, \( N(t) \sim P_\alpha(\lambda) \) the duration between events follows an exponential distribution of the parameter \( \lambda: d_i \sim \exp(\lambda), \quad i=1,2,... \) and \( d_i \) is independent. \( \lambda \) is the constant hazard function. The particularity of BIN models is that \( \lambda \) is random.

**Bin model and its properties**

**Structure of the model**

In order to model the sequence \{ \( N_i \) \} Ridberg and Shephard (2000) suggested that the BIN models specify one-step ahead forecast distribution of \{ \( N_i \) \} series using a counting distribution. In particular, they specify \( N_i | F_{i\Delta} \sim P_\alpha(\lambda_i)\) allowing \( \lambda_i \) depending upon \( F_{i\Delta} \) information available infinitesimally before time \( i\Delta \). Here \( P_\alpha(\lambda_i) \) denotes a Poisson distribution with mean \( \lambda_i \), where \( \lambda_i \) is a linear function of past data as moving average models: BIN (1,1). Then the BIN(1,1) is given as follows:

\[ N_i | F_{i\Delta} \sim P_\alpha(\lambda_i), \quad \lambda_i = \alpha + \gamma N_{i-1} + \delta i_{i-1}, \quad (3) \]

which is labeled a BIN(1,1) model. Sufficient conditions for \( \lambda_i \) to be non-negative is that \( \alpha, \gamma, \delta \geq 0 \). This model is inspired by the GARCH model according to Bollerslev (1986) and Taylor (1986). This model is thus an autoregressive moving average (ARMA) type for:

\[ N_i = \lambda_i + u_i \quad (4) \]

\[ = \alpha + \gamma N_{i-1} + \delta i_{i-1} + u_i \]

\[ = \alpha + \gamma N_{i-1} + \delta (N_{i-1} - u_{i-1}) + u_i \quad (5) \]

\[ = \alpha + (\gamma + \delta) N_{i-1} + u_i - \delta u_{i-1}, \quad (6) \]

which can be analyzed as standard multivariate ARMA models with white noise error term, where \( u_i = N_i - \lambda_i \) such that \( E(u_i | F_{i\Delta}) = 0. \quad (3) \) Then \( u_i \) is conditionally independent and identically distributed (\( u_i \sim c.i.i.d \)).

\(^3\) \( u_i \) is consider as a Martingale since \( \lambda_i \) is the compensatory of \( N_i \).
Many of the interesting features of the BIN model follow from this structure. \( u_i = N_i - \lambda_i \) is a martingale difference (MD) which appears as an innovation for \( N_i \). This equation (as in the case of ACD(1,1)) shows that a BIN(1,1) process corresponds to a constrained ARMA(1,1) representation for \( N_i \), with an autoregressive coefficient \( \gamma + \delta \), and moving average coefficient \( -\delta \), and with a MD error term if \( \gamma + \delta < 1 \). The autocorrelation function (ACF) could be obtained by the standard formulae for the ARMA(1,1) model. The main features (mean, variance, autocorrelation function) of this model are subsequently described.

**Statistical properties of the BIN models**

By definition, the conditional expectation of \( \{N_i\} \) is equal to \( \lambda_i \). Then, equation (3) allows us to forecast expected counts based on the information set at the previous period. If \( \{N_i\} \) is generated from equation 3 for \( n \in \mathbb{Z} \) with \( \alpha, \gamma, \delta > 0 \), then \( \gamma + \delta < 1 \), the unconditional expectation (\( \mu \)) and variance (\( \sigma^2 \)) of \( \{N_i\} \) are given by:

\[
\mu = E(N_i) = \frac{\alpha}{1-(\delta + \gamma)} , \quad (7)
\]

\[
\sigma^2 = \frac{\mu (1-\delta^2 -2\gamma\delta)}{1-(\delta + \gamma)^2} = \mu + \frac{\gamma^2}{1-(\delta + \gamma)^2} , \quad (8)
\]

And the autocorrelation function is derived as:

\[
\rho_1 = \frac{\gamma(1-\delta(\gamma + \delta))}{1+\delta^2 -2\delta(\gamma + \delta)}, \quad \rho_s = \rho_1(\gamma +\delta)^{s-1},
\]

\( s = 2,3,... \)

Another way to compute \( \rho_s \) is \( \rho_s = \rho_{s-1}(\gamma + \delta) \) \( (9) \).

From the expression of \( (\sigma^2) \), it is easy to check that \( (\sigma^2) \) measures overdispersion is greater than zero. \( N_i \) becomes more dispersed than \( u_i \) when \( \gamma \) increases. From equation (8), it is easy to check that \( \sigma^2 \) is greater than \( \mu \) if \( \gamma \) is greater than zero, and implies that the series observations is overdispersed.

The properties of \( \lambda_i \) are sometimes helpful, in particular \( E(\lambda_i) = \mu \) and \( Var(\lambda_i) = \sigma^2 - \mu \)

\[
= \mu \frac{\gamma^2}{1-(\delta + \gamma)^2} = \frac{\alpha \gamma^2}{\{1-(\delta + \gamma)\}\{1-(\delta + \gamma)^2\}};
\]

\[
Cov(N_i, \lambda_i) = E(\lambda_i^2) - \{E(\lambda_i)\} = Var(\lambda_i)
\]

**Generalization of BIN(1,1) model**

BIN (1,1) model can be generalized as a BIN model of order \( p,q \) (BIN (p,q)) where

\[
N_i | F_{i-1} \sim P_0(\lambda_i),
\]

\[
\lambda_i = \alpha + \sum_{j=1}^{p} \gamma_j N_{i-j} + \sum_{j=1}^{q} \delta_j \lambda_{i-j}.
\]

As in the standard ARMA case, \( p \) denotes the number of autoregressive terms in the model and \( q \) the number of moving average.

We have: \( \alpha > 0, \gamma_j > 0, \delta_j > 0 \) that leads to the fact that \( \lambda_i \) is a non-negative sequence when \( q \geq 1 \) with a probability of 1. The ARMA representation of this model can be written as follows:

\[
N_i = \lambda_i + u_i = \alpha + \sum_{j=1}^{p} \gamma_j N_{i-j} + \sum_{j=1}^{q} \delta_j (N_{i-j} - \lambda_{i-j}) + u_i = \alpha + \sum_{j=1}^{\max(p,q)} \phi_j N_{i-j} + u_i - \sum_{j=1}^{q} \delta_j u_{i-j},
\]

where \( u_i = N_i - \lambda_i \) is also a martingale difference sequence and \( \phi_j = \gamma_j + \delta_j \). This model is covariance stationary when \( \sum_{j=1}^{\max(p,q)} \phi_j < 1 \).

This last assumption allows us to write the following properties
X represent the time (t) and y represent observations of events i.e. count data, according to simulation results.

\[
\mu = \frac{\alpha}{\max(p,q)} - \sum_{j=1}^{\phi} \phi_j
\]

\[
Cov(N, \lambda_i) = Var(\lambda_i) = \sigma^2 - \mu
\]

And

\[
Var(u_i) = Var(N) + Var(\lambda_i) - 2Cov(N, \lambda_i)
\]

Then, one can compute easily the autocorrelation function of \{N_i\} by using results on variances and autocovariances of ARMA \{\max(p,q), q\} process as in the case of BIN(1,1) form. The relevant work is focus on BIN(1,1) model.

**Numerical illustrations**

The first and second unconditional moments, and the autocovariances can be computed analytically as shown above. It is therefore necessary to give numerical results about these moments and autocovariances for several sets of parameters. By using (7), (8), and (9), we compute the degree of overdispersion and obtain Figure 1 which shows the autocorrelation function for four set parameters\(^4\). Similarly as in the case for the ARCH, GARCH, and ACD class of models, \(\delta\) \(\approx 1\) implies a slowly decreasing autocorrelation function, and a large value of \(\gamma\) implies a large degree of overdispersion. Figure 1 shows the theoretical and empirical autocorrelation function.

Figure 1c seems to exhibit the best-fitted representation of the model according to the residuals. Thus, the real values of parameters \(\gamma\) and \(\delta\) is close to 0.10 and 0.85, respectively. The missing parameter here is the time interval of the count data denoted by \(\Delta\). It will be taken into account in the case of the empirical work using the actual data.

\(^4\) The model is BIN(1,1) with the unconditional mean set equal to one, i.e. \(\alpha = 1 - \gamma - \delta\)
Table 1. Overdispersion of BIN model/

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Overdispersion ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.75</td>
</tr>
<tr>
<td>0.20</td>
<td>0.75</td>
</tr>
<tr>
<td>0.10</td>
<td>0.85</td>
</tr>
<tr>
<td>0.30</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Source: Estimations of parameters using DGP data.

On the other hand, we can examine the experimental overdispersion implied by changes in values of gamma and delta in the Table 1. The results are derived from a simulated data that allows a Data Generating Process. The overdispersion ratio is defined as the ratio of standard deviation / mean, computed according to the formula (7) and (8). We parameterize lambda to one, thus \( \alpha = 1 - \gamma - \delta \). In brackets we have the theoretical overdispersion ratio. Table 1 shows that the overdispersion ratio is an increasing function of gamma and decreasing function of delta in BIN(1,1) model.

Estimation by likelihood method

Consider \( N_1, \ldots, N_T \) to be the \( T \) non-negative integer events count observations for the dependent variable, that is a random dependent variable which represents the number of events (financial: quote, price or volume) that have occurred during the observation period \( t \). Let's assume that the events which occur within each period are independent and have constant rate of occurrence. Then, \( N_i \) can by this fact follow a Poisson distribution with conditional probability density function thus:

\[
f_p(N_i, \lambda_i) = \begin{cases} 
\frac{e^{-\lambda_i} (\lambda_i)^{N_i}}{N_i!} & \text{for } \lambda_i > 0 \text{ and } N_i = 0,1,... \\
0 & \text{otherwise}
\end{cases}
\]

with expected value and variance \( \lambda_i \) (the rate of event occurrence, that must be greater than zero) is assumed to be an exponential linear function of a vector of explanatory variables \( x_i \):

\[
E(N_i) = \exp(x_i, \beta)
\]

A constant term as the first element of \( x_i \) is included in the program, and one can include any number of explanatory variables.

The Poisson regression model that is the standard model for count data, is a non linear regression. This regression model is hence based upon the Poisson distribution with intensity parameter \( \lambda \) that depends on covariates regressors. In the case of missing of stochastic variation and with exact parametric dependence with exogenous covariates, then we have the standard Poisson regression. The mixed Poisson regression is obtained if the function relating \( \lambda \) and the covariates is stochastic, likely because it involves unobserved random variables thus assumptions must be made to take into account the random term for obtaining the precise form or to come back to the standard Poisson model.

The appropriate data are cross-sectional for applied work which consists of \( T \) independent observations, indexed by \( i \) \( (N_i, x_i) \). \( N_i \) is the number of occurrence of the event object of study and \( x_i \) is the vector of linearly independent regressors that are thought to determine \( N_i \). A regression model is thus based on this conditional distribution with a \( k \)-dimensional vector of covariates, \( x'_i = (x_{i1}, \ldots, x_{ik}) \), and parameters \( \beta \) through a continuous function \( \lambda(x_i, \beta) \), such that \( E[N_i|x_i] = \lambda(x_i, \beta) \). That is to say \( N_i \) given \( x_i \) is Poisson-distributed with density

\[
f(N_i|x_i) = \frac{e^{-\lambda_i} \lambda_i^{N_i}}{N_i!}, \quad N_i = 0,1,2,\ldots
\]

(1)

The log-linear form is the parameterization such that

\[
\lambda_i = \exp(x'_i \beta)
\]

to keep \( \lambda > 0 \).

The Poisson distribution property allows us to write \( V(n_i|x_i) = E(n_i|x_i) \) with \( n_i \) considered as the realization of random variable \( N_i \); then,

\[
E(n_i|x_i) = \exp(x'_i \beta)
\]

\[
= \exp(x_{i1} \beta_1) \exp(x_{i2} \beta_2) \cdots \exp(x_{ik} \beta_k).
\]

In the likelihood-based models, the joint density of the dependent variables is specified. We assume that the scalar random variable \( N_i \) given the vector of regressors \( x_i \) and parameter vector \( \theta \), is distributed with density

\[
5\text{ Here } X_i \text{ contains autoregressive components } (\lambda_{i-1}, N_{i-1}) \text{ in the case of } BIN(1,1) \text{ model.}
\]
The likelihood principle chooses as estimator of $\theta$ the value that maximizes the joint probability of observing the sample values $N_1, \ldots, N_T$. This probability is called the likelihood function, and it appears as a function of parameters conditional on the data. It is denoted as:

$$L(\theta) = \prod_{i=1}^{T} f(n_i|x_i, \theta),$$

(3)

This formulation allows suppressing the dependence of $L(\theta)$ on the data and has assumed independence over $i$. This definition could be extended to time series data by allowing $x_i$ to include lagged dependence and independent variables, even if it implicitly assumes cross-section data.

Therefore maximizing the likelihood function is equivalent to maximizing the log-likelihood function

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{T} \ln f(n_i|x_i, \theta)$$

(4)

Under the so-called regularity conditions that are conditions on continuity and differentiation, the Maximum Likelihood Estimator (MLE) $\hat{\theta}_{ML}$ is the solution to the first order conditions.

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^{T} \frac{\partial \ln f_i}{\partial \theta} = 0$$

(5)

where $f_i = f(n_i|x_i, \theta)$ and $\frac{\partial l}{\partial \theta}$ is a $q \times 1$ vector.

The data generating process for $n_i$ has density $f(n_i|x_i, \theta_0)$ where $\theta_0$ is the true parameter value. That is, to say the asymptotic distribution of the MLE is usually obtained under the assumption that the density is correctly specified. Then, under the regularity conditions, $\hat{\theta} \Rightarrow \theta_0$, the MLE is consistent for $\theta_0$.

Then,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \stackrel{d}{\Rightarrow} N[0, A^{-1}].$$

(6)

where the $q \times q$ matrix $A$ defined as

$$A = -\lim_{T \to \infty} \frac{1}{T} E \left[ \sum_{i=1}^{T} \frac{\partial^2 \ln f_i}{\partial \theta \partial \theta'} | \theta_0 \right]$$

(7)

Simulation

In this section, we perform an algorithm for reference sample generation and then we compute the characteristics of the theoretical model (after a parameterization), and the parameters from the generated sample. We did comparison between the two generated models. We also compute the autocorrelation function (Figure 1).

By Monte Carlo simulation of discrete distribution (Rubinstein, 1981), we can perform the estimation of Poisson distribution, but there exists in Gauss the command that allows the estimation by Poisson process.

Comparison analysis

Data Generating Process (DGP) allows creating a sample of 1000 observations. The DGP is a method that allows to create artificial data by using a function and the techniques of parameterization. Therefore, one can use a spread of numbers to compute different values of the function. These data could be used to do different fitted regressions according to coefficients, therefore it will be possible to choose the best fitted regression.

The simulation allows us, by a play on parameters, to obtain the following experimental results in Table 2.

We parameterize $\lambda$ to 1, thus $\alpha = 1 - \gamma - \delta$ in all computations.

The results allow us to draw the following conclusion: $\lambda = 1.00$, $\alpha = 0.05$, $\gamma = 0.10$, $\delta = 0.85$ (having the best fitted model). Then, the means of theoretical and empirical results are 1.00 and 1.099, respectively. The standard deviations of the theoretical and empirical results are 1.050 and 1.041, respectively. The theoretical and empirical overdispersions are 1.050 and 1.032, respectively and are the lower overdispersion values for theoretical and empirical results. The graph of the autocorrelation function in this case indicates that it is the best-fitted model.

It is easy to check that the overdispersion coefficient, which is equal to, $\frac{\sigma^2}{\mu^2}$ increases with $\gamma$.

Other tools to check the adequacy of the model are skewness and kurtosis. The skewness measures the locations by indicating the number around which the sample data are centered. It indicates the direction in which a frequency distribution (or frequency curve or frequency polygon) leans. Thus, skewness equal to zero implies a symmetric distribution. We can also have the case of negative or positive skewness. Kurtosis measures a distribution’s peakedness that is, the degree to which one narrow range of values contains a large fraction of sample...

$^6 \theta = \begin{pmatrix} \alpha \\ \gamma \\ \delta \end{pmatrix}$ in the BIN(1,1) case.
Table 2. Statistics of simulated data

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ = 0.10 γ̃</td>
<td>0.75</td>
<td>1.021 (1.000)</td>
<td>1.050</td>
<td>4.153</td>
</tr>
<tr>
<td>γ = 0.20 γ̃</td>
<td>0.75</td>
<td>0.976 (1.000)</td>
<td>1.494</td>
<td>5.882</td>
</tr>
<tr>
<td>γ = 0.10 γ̃</td>
<td>0.85</td>
<td>1.009 (1.000)</td>
<td>1.101</td>
<td>4.384</td>
</tr>
<tr>
<td>γ = 0.30 γ̃</td>
<td>0.65</td>
<td>0.920 (1.000)</td>
<td>1.813</td>
<td>7.317</td>
</tr>
</tbody>
</table>

Source: Estimations of parameters using DGP data.

Table 3. Estimation results by MLE using DGP count data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0.143 (0.150)</td>
<td>3.071</td>
</tr>
<tr>
<td>γ</td>
<td>0.109 (0.100)</td>
<td>4.743</td>
</tr>
<tr>
<td>Δ</td>
<td>0.752 (0.750)</td>
<td>12.981</td>
</tr>
<tr>
<td>Q(10)</td>
<td>79.87</td>
<td></td>
</tr>
<tr>
<td>Q(10)*</td>
<td>9.77</td>
<td></td>
</tr>
</tbody>
</table>

Source: estimations of parameters using DGP count data.

Skewness indicates whether the histogram "leans to the left (negative value)" or "leans to the right (positive value)", while kurtosis indicates how peaked it is. Their formulas are

\[ \text{Skewness, } sk = \frac{M^3}{M^2^3}, \quad \text{Kurtosis, } kur = \frac{M^4}{M^2^2}, \]

where \( M_k \) is the \( k \)th moment of a sample of ungrouped data.

The simulation results give \( sk = 1.10 \), showing that the frequency curve (or density curve) leans to the right. This indicates that there are more values to the right of the model than to the left; while \( kur = 4.38 > 3 \) implies a peak curve. These results conform to Poisson density function.

Validation of the model: Monte Carlo method

For this purpose, we use the data obtained by DGP for Maximum Likelihood estimation, and the results are shown in Table 3, 4, 5, 6. Q(10) and Q(10)* correspond respectively to the Ljung-Box Q-statistic of order 10 on counts \( N_i \) and Q-statistic on the residual \( u_i \) defined in the BIN(1,1) model. If Q is more than 18.307, then there is autocorrelation of order 10 for a threshold of 5%. If the t-stat is greater than 1.96 then the parameter is significant at a threshold of 5%. The results in Table 3 and Figure 2 show that the BIN(1,1) model fits properly and may be used for estimation and forecasting.

Application to NYSE data

Data of three stocks

For empirical analysis, we choose a financial activity of three stocks traded on the NYSE: BOEING, DISNEY and AWK (American Water Work). Data previously used for ACD by Bauwens, Giot and Veredas, were extracted from the Trade and Quote (TAQ) database of the NYSE (Bauwens and Giot, 1999). For the three stocks, we choose quote data for BOEING, quotes volume for DISNEY and trades for AWK. Before using these data, we must transform durations to
Figure 2. Counts and forecast counts for estimation of Table 3

Table 4. Estimation results by MLE using DGP count data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.068 (0.050)</td>
<td>4.546</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.244 (0.200)</td>
<td>9.963</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>0.694 (0.750)</td>
<td>22.397</td>
</tr>
<tr>
<td>$Q(10)$</td>
<td>1416.44</td>
<td></td>
</tr>
<tr>
<td>$Q(10)^*$</td>
<td>13.97</td>
<td></td>
</tr>
</tbody>
</table>

Source: estimations of parameters using DGP count data.

Figure 3. Counts and forecast counts for estimation of Table 4
Table 5. Estimation results by MLE using DGP count data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0.038 (0.050)</td>
<td>2.464</td>
</tr>
<tr>
<td>γ</td>
<td>0.085 (0.100)</td>
<td>5.312</td>
</tr>
<tr>
<td>Δ</td>
<td>0.88 (0.850)</td>
<td>34.996</td>
</tr>
</tbody>
</table>

Q(10) = 159.88
Q(10)* = 15.24

Source: Estimations of parameters using DGP count data.

Figure 4. Counts and forecast counts for estimation of Table 5.

Table 6. Estimation results by MLE using quote count data for BOEING stock

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ</td>
<td>0.25</td>
</tr>
<tr>
<td>α</td>
<td>0.009 (0.002)</td>
</tr>
<tr>
<td>γ</td>
<td>0.054 (0.006)</td>
</tr>
<tr>
<td>δ</td>
<td>0.909 (0.014)</td>
</tr>
<tr>
<td>Q(10)</td>
<td>829.49</td>
</tr>
<tr>
<td>Q(10)*</td>
<td>41.08</td>
</tr>
</tbody>
</table>

Source: Estimations of parameters using quote count data for BOEING stock.
In brackets we have the standard deviation. Q(10) and Q(10)* are the Ljung-Box Q-statistic of order 10 on counts \( N_i \) and Q-statistic on the residual \( u_i \) defined in the BIN(1,1) model. Δ is the fixed length of time.

We obtain the data under the count data form according to Veredas Program, which transforms durations to counts. The count is made for a fixed length of time. In our case we use an interval of 0.25 second, 1 second and 4 seconds and estimations are performed using each form of data. The estimation method used here is the Maximum Likelihood method.

In brackets we have the standard deviation. Q(10) and Q(10)* are the Ljung-Box Q-statistic of order 10 on counts \( N_i \) and Q-statistic on the residual \( u_i \) defined in the
Table 7. Estimation results by MLE using quote volume count data for DISNEY stock.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$</td>
<td>0.25</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.002 (0.000)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.009 (0.001)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.984 (0.002)</td>
</tr>
</tbody>
</table>

Q(10)  | 375.60 |
Q(10)* | 410.29 |

Source: Estimations of parameters using quote count data for DISNEY stock. In brackets we have the standard deviation. Q(10) and Q(10)* are the Ljung-Box Q-statistic of order 10 on counts ($N_i$) and Q-statistic on the residual $u_i$ defined in the BIN(1,1) model. $\Delta$ is the fixed length of time.

Table 8. Estimation results by MLE using trades count data for AWK stock

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$</td>
<td>0.25</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.004 (0.001)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.012 (0.002)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.973 (0.006)</td>
</tr>
</tbody>
</table>

Q(10)  | 147.74 |
Q(10)* | 26.41  |

Source: Estimations of parameters using quote count data for AWK stock. In brackets we have the standard deviation. Q(10) and Q(10)* are the Ljung-Box Q-statistic of order 10 on counts ($N_i$) and Q-statistic on the residual $u_i$ defined in the BIN(1,1) model. $\Delta$ is the fixed length of time.

Conclusion

The aim of this survey is to build and test BIN (1,1) model for the counts data. Before generating the data by parametrization, we used these data for estimation by the ML method for BIN(1,1) validation. The results exhibit a good behaviour of this model and so can be applied to actual data. For application, we use the transformed data (durations to counts) at different fixed length of time. The results of estimation of the three stocks (BOEING, DISNEY, and AWK) by the Maximum Likelihood method allows for the following conclusion:

There is less dependence between high frequency data when we consider a large value at a fixed length of time ($\Delta$) as the value of gamma becomes larger which increases the dispersion. The model could be generalized to take into account other variables and it could be used for density forecasting.

ACKNOWLEDGEMENT

I wish to thank Professor BAUWENS Luc, for the
supervision of this work, Professor DEHEZ Pierre, Professor
GIOT Pierre, all to Université Catholique de Louvain
(Belgium). I am also grateful to Dr. VEREDAS David for his
precious collaboration.
Bauwens L (1999) "Recent developments in the
ecnometrics of financial markets using intra-day data." CORE discussion paper 1403, Université Catholique de
Louvain (UCL).

REFERENCE

Application To The Bid-Ask Quote Process of Three NYSE
Stocks." Annales d’Economie et de Statistique 60:117—
149.
duration model: a latent factor model for the analysis of
financial durations." J. Econ. 119:381-412.Crossref
Bera AK, Higgins ML (1993). "ARCH models: properties,
Crossref
Bertram WK (2004). An empirical investigaion of
Australian Stock Exchange data. Physica A 341:533-
546.Crossref
Exchange equities. Physica A 346 561-576.Crossref
Imbalances and Stock Price Movements on October 19
Bonanno G, Lillo F, Mantegna R (2000). Dynamics of the
number of trades of financial securities. Physica A
280:136-141.Crossref
walks. Chaos 15 026104.Crossref
Bouchaud JP, Matacz A, Potters M (2001). Leverage Effect in
Rev. Lett. 87:228701.Crossref
Cameron AC, Trivedi PK (1998), "Regression analysis of
count data."Cambridge University Press.Crossref
Chan K, Fong WM (2000). Trade size, order imbalance, and
the volatility-volume relation. J. Financ. Econ. 57:247-
273.Crossref
Chordia T, Roll R, Subrahmanyam A (2002). Order
imbalance, liquidity, and market returns. J. Financ. Econ.
65:111-130.Crossref
wins? study of long-run trader survival in an artificial
fluctuations in financial markets." Macroecon. Dyn. 4 170-
196.Crossref
Dacorogna M, Gençay R, M’uller U, Olsen RB, Pictet O
Academic Press.
Eisler Z, Kertész J (2006)."Size matters: some stylized facts
of the stock market revisited." Eur. Phys. J. B.Crossref
Extremal Events for Insurance and Finance. Springer,
Berlin.Crossref
Engle R, Russell J (1997). Forecasting the frequency of
changes in quoted foreign exchange prices with the
autoregressive conditional duration model. J. Empir.
Financ. 4:187–212.Crossref
University Press.
Engle RF, Granger CWJ (1991) "Long run economic
relationships: Readings in cointegration". Oxford
University Press.
Engle RF, Russell JR (1994). Forecasting transaction rates:
the autoregressive conditional duration model. NBER
Working paper series 4966.
Engle RF, Russell JR (1998). Autoregressive conditional
duration: a new model for irregularly spaced transaction
data. Econometrica 66:1127-1162.Crossref
Fernandes M, Gramming J (2006). A family of
autoregressive conditional duration models. J. Econ.
130:1-23.Crossref
GAUSS Applications (1995) "Constrained Maximum
Likelihood." Aptech Systems, Inc. Maple Valey, USA.
volatility duration models. J. Econometrics 119:413-
433.Crossref
Giot P(1999) "Time transformations, intra-day data and
volatility models." CORE discussion paper 9944, UCL.
Goll S (2001). Long memory effects in ultra-
high frequency data. Quaderni di Statistica 3 43-52.Goodhart,
financial markets: issues and applications. J. Empir. Financ. 4 73-
114.Crossref
Statistical properties of share volume traded in financial
Processes. Les Cashiers du CREF.
Hardle W, Kirman A (1995). Neoclassical demand - A
model-free examination of price-quantity relations in the
Marseilles fish market. J. Econ. 67:227-257.
Trading Units on Stock Value and Price Volatility. J.
Hautsch N (2012). Econometrics of Financial High-
Frequency Data. Springer, Berlin.Crossref
Hudson RL, Mandelbrot BB (2010). The (Mis)Behaviour of
Markets. Profile Business.