Exponential Multivariate Autoregressive Conditional High Frequency Data Model

Gustavo Santos Raposo
Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro

Alvaro Veiga
Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro

October 17, 2004

Draft Version

Abstract

The modeling of financial transaction data – price, spread, volume and duration – in an event basis is motivating a growing number of works. The first proposals, where limited to pure duration models. Then its impact on the volatility was analyzed. More recently a vector model also including volume was studied by Manganelli (2002). In this paper, we extend his work by including the bid-ask spread into the model throughout a vector autoregressive model. The conditional means of spread, volume and duration along with the volatility of returns evolve through transaction events based on an exponential formulation that we called Exponential Multivariate Autoregressive Conditional Model (EMACM).

In this new proposal, there is no constrains on the parameters. This facilitates the maximum likelihood estimation of the model and allows the use of simple likelihood ratio hypothesis tests to specify the model and obtain some clues about the interdependency structure of the variables.

Keywords: High frequency data, GARCH, autoregressive conditional multivariate models, nonlinear time series.

1 Introduction

The availability of high frequency databases makes possible to understand financial market dynamics (intra-day basis) and test some of hypothesis brought up by the microstructure theory. In that way, many formulations have been suggested.

Historically, we can observe three distinct phases when considering trading variables modeling. The first corresponds to the early developments made. In the second, the concepts embedded in ARCH/GARCH models, formulated in order to deal with volatility regimes in stock price returns, were applied to model other trading variables, specially the time between financial transactions (duration).

Now, in the third phase, the focus is not only on the dynamic of a specific high frequency variable, but also on the influence that exists among them. Here, the main goal is to define how the trading variables influence each other.

Regarding high frequency data models, the first development occurred in 1948 when Wold (Skandinavisk Aktuarietidskrist - *On Stationary Point Process and Markov Changes*) proposed to capture the dynamic presented in the conditional intensity through the use of ARMA models. In 1955, Cox (Journal of the Royal Statistical Society - *Some Statistical Models Connected with Series of Events*) included lagged variables in order to explain the conditional intensity. Latter, in 1980, Lewis (Advances in Applied Probability - *First-Order Autoregressive Gamma Sequences and Point Processes*) extended the original proposal of Cox - EARMA.

In 1998, Engle e Russell (Econometrica - *Autoregressive Conditional Duration: A New Model For Irregularly Spaced Transaction Data*) introduced the ACD (*Autoregressive Conditional Duration*) model, in which the time between events has been described as a sequence of independent random variables with a time varying mean given by a GARCH type equation. Bauwens and Veredas, in 1999, defined a stochastic process for conditional duration (latent stochastic factor), in order to capture the market information flow (non-observable variable). Later, Bauwens and Giot (2000) proposed the use of a logarithmic version of ACD models, where the non-negativity constraint wasn't necessary.

In 2000, Engle incorporated the methodology developed before, in a volatility context (UHF-GARCH). After that, in 2001, Zhang, Russell and Tsay (Econometrica – *A nonlinear autoregressive conditional duration model with applications to financial transaction data*) extended the original model (EACD and WACD), through the using of thresholds (multiple regimes), in order to capture the non-linearity. In a similar way, in 2001, Fernandes and Gramming (CORE – *A family of autoregressive conditional duration models*) added some changes to the model initially proposed by Engle and Russell, dealing with non-linearity through the application of a Box-Cox transformation over the original series.

Recently, in 2002, Manganelli (ECB Working Paper Series – *Duration, Volume and Volatility impact of trades*) proposed the joint modeling of different variables (duration, volume and volatility) involved in the financial transaction process by the use of Vector Autoregressive Models (VARM). That was the first time in which volume information has been explicitly modeled and the work is one of the first trials of dealing with joint behavior of trading variables.

In this paper, we extend the work of Manganelli by including the bid-ask spread into an autoregressive multivariate system and proposing an exponential formulation to the conditional mean, avoiding the adoption of constraints in the parameters when maximizing the likelihood function. We called it the Exponential Multivariate Autoregressive Conditional Model (EMACM). The structure of the coefficient matrices of the system is tested via likelihood ratio tests, answering some of the questions raised in the microstructure literature about causality and dependency among variables.

Regarding the analysis of intra-day seasonal pattern, we've found that the lowest durations (high intensity) are observed close to the opening and closing of financial market. As a consequence, the bid-ask spread and price volatility increase. In relation to the volume intra-day pattern, the highest values are observed close to the opening of transaction days, what can be explained by the fact that new information were not incorporated into price (after-market effects).

Considering the structure of the coefficient matrices, the likelihood ratio test pointed to the rejection of the hypothesis of no causality in trading day variables, since the individual formulation is strongly rejected. Here, the results show that the system seems to be variation-free as suggested by Manganelli.

That article is divided as follows. Section 2 presents the model. Section 3 describes the seasonal adjustment (off-line estimation). Section 4 brings details of the nonlinear optimization algorithms used. A Monte-Carlo simulation is carried out in Section 5 and an empirical example is shown in Section 6. Finally, Section 7 concludes.

2 The Model

The model proposed in this paper is called Exponential Multivariate Autoregressive Conditional High Frequency Data Model (EMACM).

Lets define x_i as being the duration of the i-th observed financial transaction, where $x_i = t_i - t_{i-1}$ (time space between trades) and z_i as a vector of explanatory variables. Thus:

$$(x_i, z_i) \sim f(x_i, z_i | \Omega_i; \theta) \tag{2.1}$$

Where, $f(x_i, z_i | \Omega_i; \theta)$ corresponds to the joint probability distribution function of x_i and z_i, Ω_i is the remaining information available until the i-th event has occurred and θ is the vector of unknown parameters.

Letting $z_i' = (v_i \ s_i \ y_i)$, where: v_i is the volume of i-th transaction, s_i is the bid-ask spread and y_i is instantaneous return. Thus,

$$(x_i, z_i) \sim f(x_i, v_i, s_i, y_i | \Omega_i; \theta) \tag{2.2}$$

Re-writing the joint probability distribution function (2.2) as the product of the conditional probability distribution, we have:

$$(x_i, z_i) = g(x_i | \Omega_i; \theta_1) h(v_i | x_i, \Omega_i; \theta_2) k(s_i | x_i, v_i, \Omega_i; \theta_3) l(y_i | x_i, v_i, s_i, \Omega_i; \theta_4) \tag{2.3}$$

Equation 2.2 seems natural when the use of strategic models is considered. For example, Kyle (1985) modeled the informed traders behavior based on the effect of buy or sell orders into price, conditioning the analysis to the non-informed traders and market-makers attitude.

At the moment the information became public, it's verified a strong offer/demand pressure originated, mainly, by the action of market makers. Once the time interval among events and the volume could indicate that some traders may be using private information, the market-makers will use that in order to prevent losses. In that way, investors who own some information that is not disseminated in market will split their trades (decreasing volumes per transaction and increasing the number of transactions per time unit – intensity), making the identification process more complicated, postponing any changes in the bid and ask prices and, consequently, in the transaction price itself.

Based on what was disposed by Kyle, it's clear the option made in favor of the relation establish by equation 2.3, in which the causality relation is shown in figure 1.

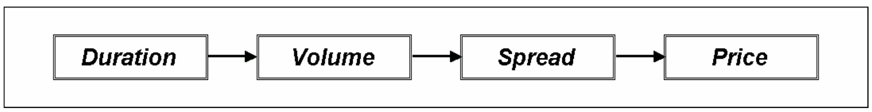


Figure 1 – Causality relation (duration, volume, spread and price)

Defining each one of the system components, the models can be determined separately:

- Duration:

$$x_i = \psi_i \,.\, \varepsilon_i \quad \rightarrow \quad \varepsilon_i \sim \exp(1) \tag{2.4}$$

$$\psi_i = E\left(x_i \middle| \, \Omega_i ; \theta_x\right) \tag{2.5}$$

    - Duration: positive real numbers;

- Volume: analogous to the duration models.

$$v_i = \phi_i \,.\, \eta_i \quad \rightarrow \quad \eta_i \sim \exp(1) \tag{2.6}$$

$$\phi_i = E\left(\upsilon_i \middle| \, \Omega_i ; \theta_\upsilon\right) \tag{2.7}$$

    - Volume: positive real numbers;

- Bid-ask spread: analogous to the duration models.

$$s_i = P_{sell} - P_{buy} \tag{2.8}$$

Where, P_{buy} corresponds to the buy price offered by the market makers (P_{sell} is analogous).

$$s_i = \varphi_i \,.\, \varpi_i \quad \rightarrow \quad \varpi_i \sim \exp(1) \tag{2.9}$$

$$\varphi_i = E\left(s_i \middle| \, \Omega_i ; \theta_s\right) \tag{2.10}$$

- Spread: positive real numbers;

- GARCH:

$$y_i = \sigma_i \, . \, \varsigma_i \quad \rightarrow \quad \varsigma_i \sim N(0,1) \tag{2.11}$$

$$\sigma_i^2 = E\left(y_i^2 \middle| \, \Omega_i ; \theta_y \right) \tag{2.12}$$

- Volatility: positive real numbers;

Thus, the conditional mean of the Exponential Multivariate Autoregressive Conditional High Frequency Data Model being considered could be defined as follows:

$$\ln(\mu_i) = \gamma + \sum_{k=1}^{q} A_k \ln(\mu_{i-k}) + \sum_{m=0}^{p} B_m \ln(\tau_{i-m}) \tag{2.13}$$

Where, $\mu_i' = (\psi_i \; \phi_i \; \varphi_i \; \sigma^2_i)$, $\tau_i' = (d_i \; v_i \; s_i \; y^2_i)$, γ is the vector of coefficients and $A_1, ..., A_q$ and $B_0, ..., B_p$ are matrices of coefficients of each one of the stochastic processes of the system.

The general formulation of the complete model can be written as follows.

- **Observation equation:**

$$\begin{bmatrix} x_i \\ \upsilon_i \\ s_i \\ y_i \end{bmatrix} = \begin{bmatrix} \psi_i & 0 & 0 & 0 \\ 0 & \phi_i & 0 & 0 \\ 0 & 0 & \varphi_i & 0 \\ 0 & 0 & 0 & \sigma_i \end{bmatrix} . \begin{bmatrix} \varepsilon_i \\ \eta_i \\ \varpi_i \\ \varsigma_i \end{bmatrix} \tag{2.14}$$

Where, ε_i, η_i e $\omega_i \sim \exp(1)$ e $\zeta_i \sim N(0,1)$.

- **State equation:**

$$\begin{aligned} \ln\begin{bmatrix} \psi_i \\ \phi_i \\ \varphi_i \\ \sigma_i^2 \end{bmatrix} &= \begin{bmatrix} a_0 \\ b_0 \\ c_0 \\ d_0 \end{bmatrix} + \sum_{l=1}^{q} \begin{bmatrix} a_1^{(l)} & a_2^{(l)} & a_3^{(l)} & a_4^{(l)} \\ b_1^{(l)} & b_2^{(l)} & b_3^{(l)} & b_4^{(l)} \\ c_1^{(l)} & c_2^{(l)} & c_3^{(l)} & c_4^{(l)} \\ d_1^{(l)} & d_2^{(l)} & d_3^{(l)} & d_4^{(l)} \end{bmatrix} \ln\begin{bmatrix} \psi_{i-l} \\ \phi_{i-l} \\ \varphi_{i-l} \\ \sigma_{i-l}^2 \end{bmatrix} \\ &+ \begin{bmatrix} 0 & 0 & 0 & 0 \\ b_5 & 0 & 0 & 0 \\ c_5 & c_6 & 0 & 0 \\ d_5 & d_6 & d_7 & 0 \end{bmatrix} \ln\begin{bmatrix} x_i \\ \upsilon_i \\ s_i \\ y_i^2 \end{bmatrix} + \sum_{m=1}^{p} \begin{bmatrix} a_5^{(m)} & a_6^{(m)} & a_7^{(m)} & a_8^{(m)} \\ b_6^{(m)} & b_7^{(m)} & b_8^{(m)} & b_9^{(m)} \\ c_7^{(m)} & c_8^{(m)} & c_9^{(m)} & c_{10}^{(m)} \\ d_{11}^{(m)} & d_{12}^{(m)} & d_{13}^{(m)} & d_{14}^{(m)} \end{bmatrix} \ln\begin{bmatrix} x_{i-m} \\ \upsilon_{i-m} \\ s_{i-m} \\ y_{i-m}^2 \end{bmatrix} \end{aligned} \tag{2.15}$$

The system presents an iterative dynamic due to the fact that matrix B_0 is a lower triangular matrix with null main diagonal elements.

Based on equation (2.15), we can infer about the structure of the system by testing different constraints. Three structures are suggested:

- Complete model: the structure of coefficients matrices is exactly as shown in equation 2.15. In that case, the conditional mean and contemporaneous and lagged variables can influence the dynamic of the system. The causality relation is explicitly modeled.

- Variation-free: the matrices A_1, A_2, … and A_q in 2.15 are diagonal. The conditional mean of each variable cannot influence the others.

- Individual: matrix B_0 is null and the other matrices in 2.15 are diagonal (constraint model – no causality relation). Here, the conditional mean of each variable evolves based on its own lagged values and the ones of the original variable. The individual formulation is obtained (ACD, ACV, ACS and GARCH).

Equation 2.18 presents the joint likelihood function:

$$L(x, v, s, y | I_N) = \prod_{i=1}^{N}\left[\frac{1}{\psi_i}\exp\left(-\frac{x_i}{\psi_i}\right)\frac{1}{\phi_i}\exp\left(-\frac{v_i}{\phi_i}\right)\frac{1}{\varphi_i}\exp\left(-\frac{s_i}{\varphi_i}\right)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{y_i^2}{2\sigma_i^2}\right)\right] \tag{2.16}$$

Where, the conditional means ψ_i, ϕ_i, φ_i and σ_i^2 are defined by equation 2.13.

3 Intra-day Pattern Adjustment

Some of the variables being analyzed may present an intra-day pattern. As proposed by Engle and Russell (1998), all periodic or cyclic behavior should be removed before model estimation in order to avoid spurious autocorrelation. We define,

$$x_i^* = {}^{x_i}\!\big/\!_{\lambda(x_i, t_i)} \quad v_i^* = {}^{v_i}\!\big/\!_{\lambda(v_i, t_i)} \quad s_i^* = {}^{s_i}\!\big/\!_{\lambda(s_i, t_i)} \quad y_i^{2^*} = {}^{y_i^2}\!\big/\!_{\lambda(y_i^2, t_i)} \tag{3.1}$$

Where, x_i^*, v_i^*, s_i^*, y_i^{2*} are, respectively, the deseasonalized series of duration, volume, spread and volatility.

In this paper we use a natural cubic spline for each variable considered in the system equation. In order to estimate the deterministic function that will represent the different seasonal patterns, the time interval between the opening and closing of trading days was divided into equally spaced intervals of one hour. In order to increase the flexibility, one extra node was added at the end of the trading day.

Thus, the intra-day seasonal pattern is defined through the following equation:

$$\lambda(t_{i-1}) = \sum_{j=1}^{K} I_j \left[c_j + d_{1,j}(t_{i-1} - k_{j-1}) + d_{2,j}(t_{i-1} - k_{j-1})^2 + d_{3,j}(t_{i-1} - k_{j-1})^3 \right] \qquad (3.7)$$

Where,

K – number of segments;
I – variable that represents the j-th segment of the spline ($I_j = 1$ if $k_{j-1} < t_{i-j} < k_j$ and $I_j = 0$, on the contrary).

4 Estimation Process

Two different optimization algorithms were used: Quadratic Sequential Programming (SQP) for the intra-day seasonal pattern determination and the Nelder-Mead Simplex Method to system's parameters estimation, because of the discontinuities of the log-likelihood function.

4.1 Quadratic Sequential Programming:

Based on the study of Biggs (1975), Han (1977) and Powell (1978), the method should be understood as a proxy of Newton's method (nonlinear programming without constraints), for constraint problem.

Algorithm: In each iteration, the objective function Hessian is calculated through BFGS method. The Hessian is then applied in a quadratic programming sub-problem. The solution is taken as reference to the subsequent liner search procedure, initializing a new iteration.

4.2 Nelder-Mead Simplex Method:

Introduced by Nelder e Mead (1965), the main idea of the method is the determination of the minimum value point of a certain N variables function through the use of a N+1 vertices simplex.

In this method, the simplex adapts itself to the local landscape, elongating down long inclined planes, changing direction on encountering a valley at an angle, and contracting in the neighborhood of a minimum.

The criterion for stopping the process has been chosen keeping in mind the application in statistical problem involving the maximization of the likelihood function in which the unknown parameters enter nonlinearly.

Algorithm: Consider the minimization problem of a function of N variable, without constraints. Let P_0, P_1, ..., P_N as the N+1 points of the N-dimensional space defining the current simplex. Define y_i as the function value at P_i, $y_h = \max(y_i)$ for i = 0, ..., N and $y_l = \min(y_i)$ for i = 0, ..., N.

Additionally, let P_{hat} as being the centroid of the region defined by P_i's, where i is different from h and $[P_i\ P_j]$ is the distance from P_i to P_j. For each stage in the process P_h is replaced by a new point; three operations are used – reflection, contraction and expansion.

The reflection of P_h is denoted by P* and its coordinates are defined by:

$$P^* = (1+\alpha) . P_{hat} - \alpha . P_h \tag{4.2.1}$$

Where, α is a positive constant (reflection coefficient).

Thus, the point P* is on the line joining P_h and P_{hat}, on the far side of P_{hat} from P_h with [P* P_{hat}] = α . [P_h P_{hat}]. If $y_l < y^* < y_h$, so P_h will be replaced by P* and a new simplex is generated.

If $y^* < y_l$, i.e. if reflection has produced a new minimum point, so, P* is expanded to P** by the following relation:

$$P^{**} = \gamma . P^* + (1-\gamma) . P_{hat} \tag{4.2.2}$$

The expansion coefficient γ (greater than unity), corresponds to the ratio between the distances [P** P_{hat}] and [P* P_{hat}]. If $y^{**} < y_l$, P_h is replaced by P** and the process is restarted. But, if $y^{**} > y_l$, then the reflection has failed and, before restarting the process, P_h must be replaced by P*.

If, on reflecting P to P*, $y^* > y_i$, for all i different of h, so it must be defined a new P_h as being the old P_h or P* (whichever has the lower y value) and form:

$$P^{**} = \beta . P_h + (1-\beta) . P_{hat} \tag{4.2.3}$$

The contraction coefficient β lies between 0 and 1 and is the ratio between [P** P_{hat}] and [P P_{hat}]. In that way, P_h is replaced by P**, unless $y^{**} > \min(y_h, y^{**})$. If it occurs, the points P_i's are replaced by $(P_i + P_l) / 2$ and the process is restarted.

All the process finishes when the diameter of the simplex points is smaller than a pre-specified value.

Since the Nelder-Mead Method does not use the Hessian when solving the optimization problem, in order to estimate Fisher Information Matrix, it's being proposed the use of Spendley's procedure (1962). The method consists on adjusting a quadratic surface on the region composed by the N+1 simplex vertices, in the neighborhood of the optimal solution.

5 Verifying the estimation algorithm

In order to test the identification power of the proposed method, a Monte-Carlo Simulation was carried out. Here, the joint model, as described in equations 2.14 and 2.15, is simulated – EMACM(2,2). We've generated 20 realizations of the process with 1000 observations each. The parameters are estimated by maximizing the likelihood function (2.16).

Through the impulse-response function analysis the real and estimated processes are compared. The figures bellow present the main results (red dashed line – estimated impulse-response function values, red line – the mean of estimated values and blue line – impulse-response of the real data generate process - DGP).

- **Duration:**

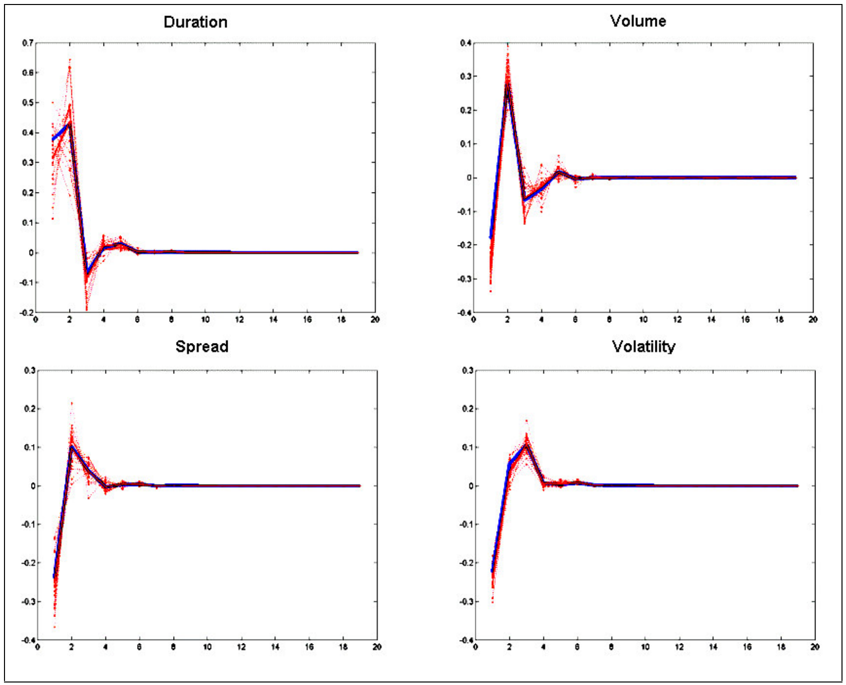


Figure 2 – Duration responses due to impulse in duration, volume, spread and volatility

- **Volume:**

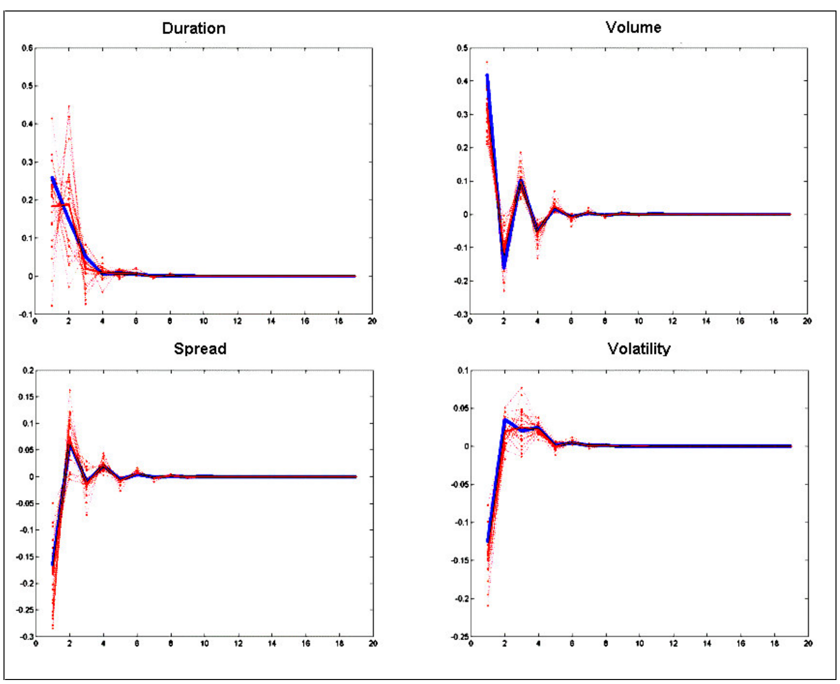


Figure 3 – Volume responses due to impulse in duration, volume, spread and volatility

- **Spread:**

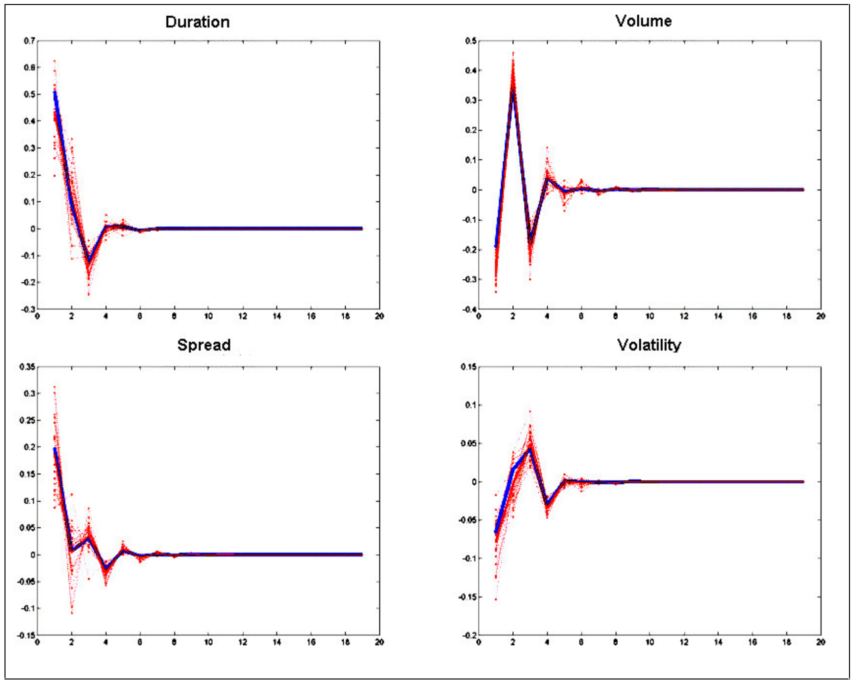


Figure 4 – Spread responses due to impulse in duration, volume, spread and volatility

- **Volatility:**

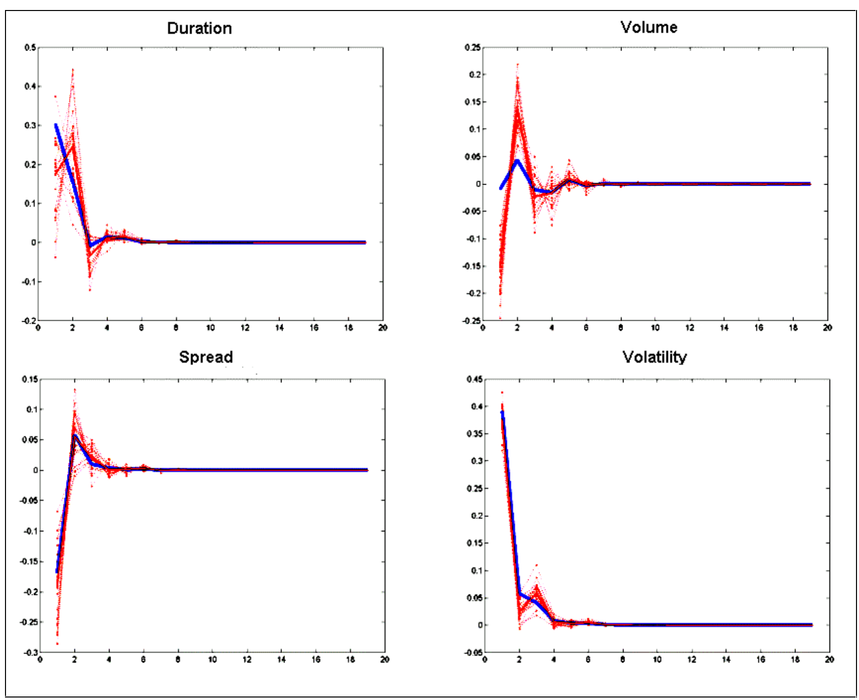


Figure 5 – Volatility responses due to impulse in duration, volume, spread and volatility

Based on the impulse-response function, we've found that the behavior of the estimated processes is close to the real one. This exercise shows that the estimation procedure presents a good identification power.

6 Empirical Analyses

6.1 Data Base:

The data used in the empirical analysis was built by Joel Hasbrouck e NYSE – Trades, Orders Reports and Quotes (TORQ). The data reflect the trades of IBM stocks, occurred between November 1st, 1990 and December 3rd, 1991.

The data comprehend all the relevant information embedded in financial transactions – buys or sells (i.e., bid price, ask price, transaction price, time and volume) registered during regular financial market time – 9:30 AM - 4:00 PM (after-market is not considered).

Since the study will focus on the modeling of tick-by-tick data (price change), some changes were implemented in the original data.

- **Duration:**
  - If the price of transaction "i" is equal to the price of transaction "i-1", the durations are added;
  - If a certain transaction presents duration equal to zero, it's removed.
- **Volume:**
  - If the price of transaction "i" is equal to the price of transaction "i-1", the volume "i" will be the mean of the volumes of both transactions.
- **Spread:**
  - If the price of transaction "i" is equal to the price of transaction "i-1", spread "i" will be the mean of the spreads "i" and "i-1" weighted by the volumes of such transactions.

Other relevant changes and considerations:

- November 23rd, 1990: was removed from the data base, due to an interruption of approximately one hour and fifteen minutes in the transactions.
- The tick-by-tick series take the unity value of the tick as reference (US$ 0.125);
- The transaction occurred during the first twenty minutes of trading day were not considered for estimation purposes (9:30 AM - 9:50 AM), because of opening postponing problems and "first trades" effects;

- For each day, the conditional mean of each one of the variables of the system (deseasonalized series) will be taken as the mean of the respective values observed between 9:50 AM and 10:00 AM (if there are no observations, the conditional mean is taken as one).

6.2 Empirical Tests:

The data set used has a total of 5806 different financial transactions. The first phase of the experiment corresponds to the estimation[1] of an EMACM(2,2), as described by equation 2.13. Here, the three structures are considered (tables I, II and III, in appendix, bring the results).

As already mentioned, before starting the estimation process the variable must be deseasonalized. Thus, following section 3, the off-line determination of the intra-day seasonal pattern is obtained. Figure 6 shows the results for each trading variable.

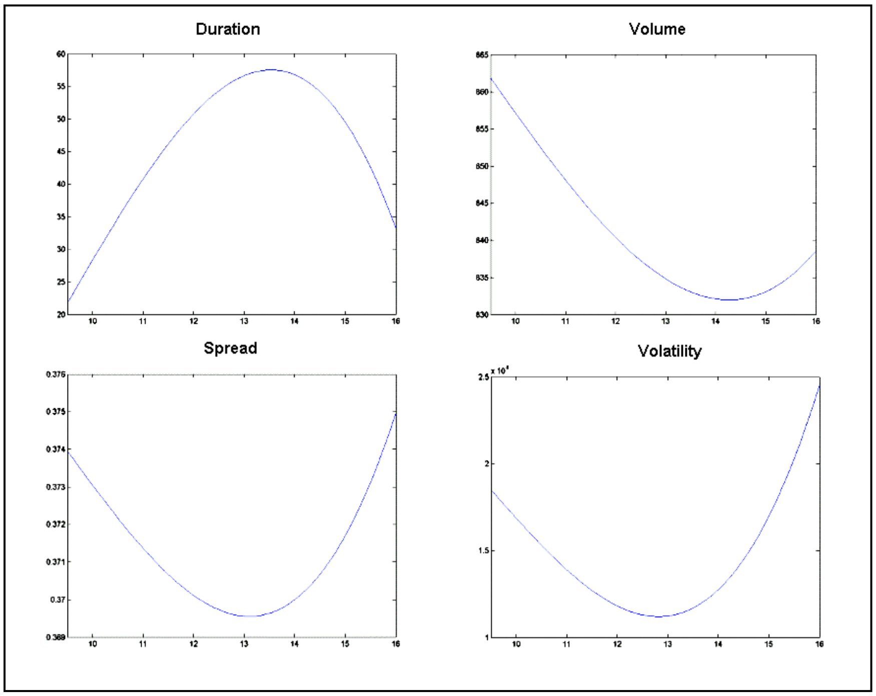


Figure 6 – Intra-day seasonal pattern

Regarding the seasonal pattern of trading variables, as pointed by Engle and Russell (1998), the highest intensity (lower durations) is observed close to the opening and closing of the trading days. Additionally, we can observe that bid-ask spread and price volatility increase, as a consequence of that fact.

Regarding volume seasonal pattern, the highest values are observed next to the opening, what can be explained by the fact that new information is not included in asset prices.

After removing seasonal effects, the system can be estimated through the use of Nelder-Mead Simplex Method. Following, the main results are presented and the three formulations proposed in that article are tested and compared.

[1] The Hessian determination is based on the study of Spendley et al (1962)

- **Complete model:**

- **ACF**

    - Duration: figure 7 shows that the model captures the linear dependence observed in original data.

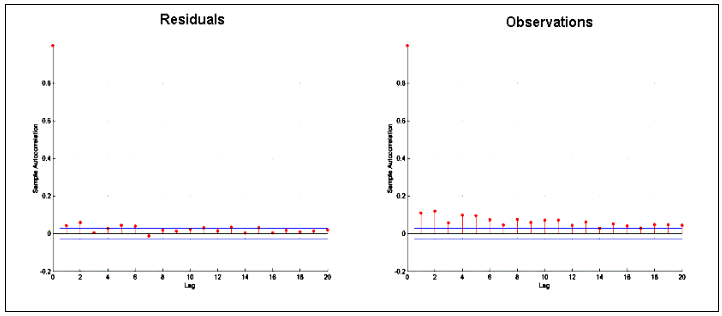


Figure 7 – ACF of duration (residuals x observations)

    - Volume: figure 8 shows that both the original data and the residuals don't present linear dependence. However, the hypothesis of null parameters in volume process has been rejected for all of them.

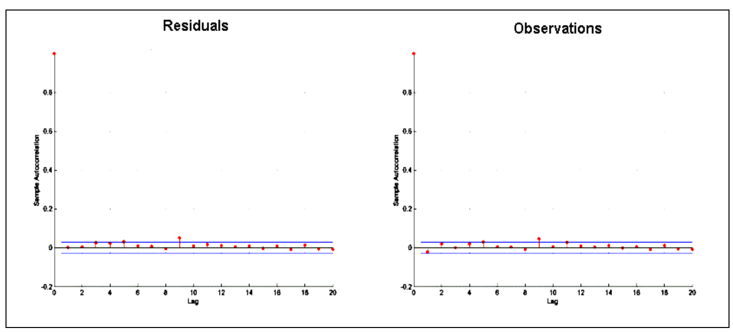


Figure 8 – ACF of volume (residuals x observations)

    - Spread: based on figure 9, we see that the model reduces the linear dependence observed in bid-ask spreads.

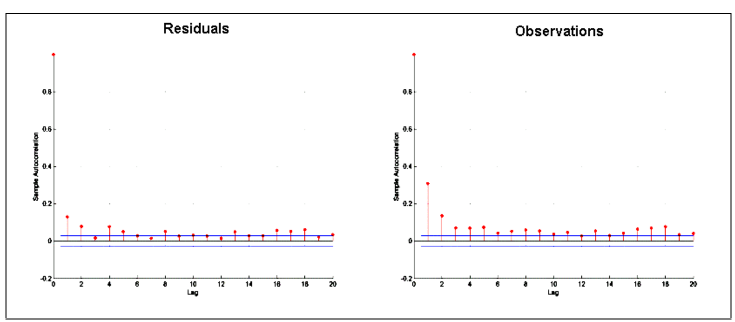


Figure 9 – ACF of spread (residuals x observations)

- Volatility: in figure 10, we observe a strong linear dependence (first lag) in instantaneous volatility (squared of instantaneous return – y^2), what is reduced but not completely captured by the model.

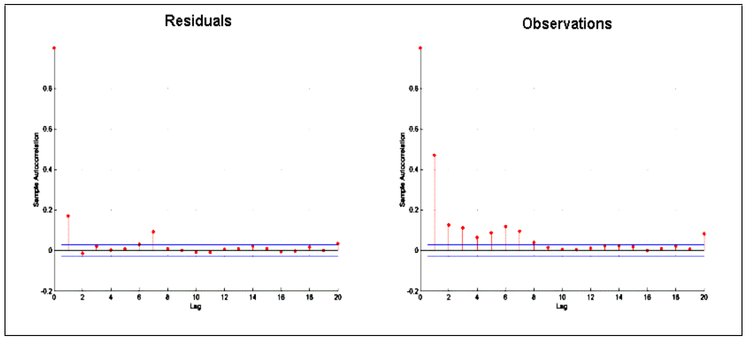


Figure 10 – ACF of volatility (residuals x observations)

- **Real x Forecasted:** figure 11 presents the trading variables plotted against the one-step-ahead forecasts.

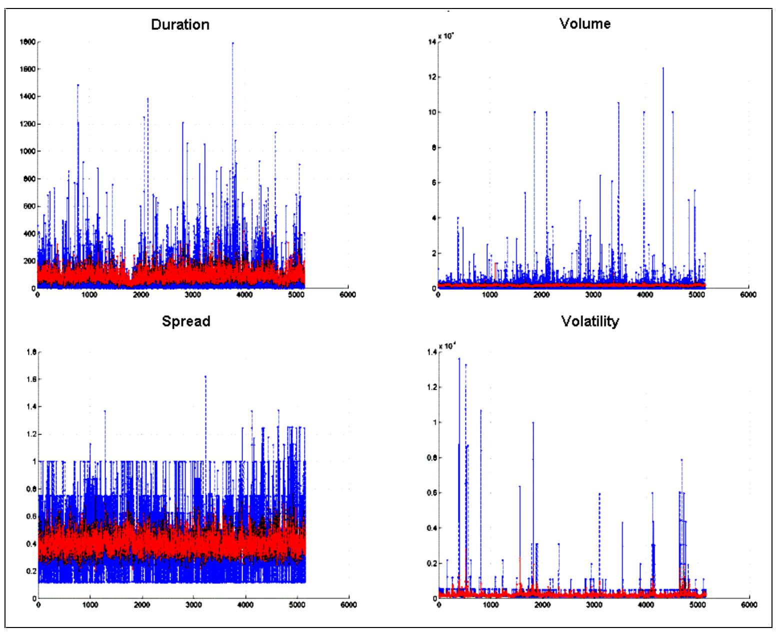


Figure 11 – Real x forecasted analysis data – model complete

- **Complete x variation-free x individual:** table 1 presents the results of Ljung-Box test. The test is based on the autocorrelation plot. However, instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags. The null hypothesis states that there are no linear dependence in data series.

		Accept H0 (95%)	P-Value	Ljung-Box	Critical Value
Observations	Duration	Reject	0,00%	442,25	25,00
	Volume	Reject	1,95%	28,34	25,00
	Spread	Reject	0,00%	774,45	25,00
	Volatility	Reject	0,00%	1486,72	25,00
Complete	Duration	Reject	0,00%	75,05	25,00
	Volume	Reject	1,30%	29,69	25,00
	Spread	Reject	0,00%	219,55	25,00
	Volatility	Reject	0,00%	206,89	25,00
Variation-Free	Duration	Reject	0,08%	38,37	25,00
	Volume	Reject	1,17%	30,04	25,00
	Spread	Reject	0,00%	295,40	25,00
	Volatility	Reject	0,00%	118,14	25,00
Individual	Duration	Reject	0,00%	40,52	25,00
	Volume	Accept	15,88%	20,35	25,00
	Spread	Reject	0,00%	419,36	25,00
	Volatility	Reject	0,00%	620,99	25,00

Table 1 – Linear dependency analysis (complete, variation-free and individual)

Based on the results obtained, it could be noticed that the proposed formulation deals quite satisfactory with the linear dependence observed in the original data. However, as seen in figure 11, there is an excessive dispersion in the residuals of all series being considered, it's probably due to the non-linearity existent in these data as observed by Engle and Russell (1998), Fernandes and Gramming (2001) and Zhang, Russell and Tsay (2001).

The constraint formulation (variation-free) has obtained a better performance when considering in-sample tracking.

In order to test the validity of constrains on the complete formulation, we use the likelihood ratio test. The null hypothesis regards the validity of the constraint formulation against the less restricted ones. Table 2 presents the results.

- **Likelihood Ratio Test:**

	Accept H_0 (95%)	P-Value	Likelihood Ratio Test	Critical Value (95%)
Complete x Variation-Free	Accept	70,06%	27,11	36,42
Complete x Individual	Reject	0,00%	1123,87	72,15
Variation-Free x Individual	Reject	0,00%	1096,76	43,77

Table 2 – Likelihood ratio test results

Considering the results of table 2, the system seems to be variation-free, as proposed by Manganelli. That hypothesis is strongly accepted (p-value = 70,06%). The hypothesis tests that consider individual formulation strongly reject the validity of independent dynamics in trading variables.

7 Conclusion and Final Comments

The EMACM is a framework to analyze high-frequency data, that allows expected duration, volume, bid-ask spread and volatility to vary according to a nonlinear function of their own lagged values. Here the exponential transformation is applied in order to guarantee the non-negativity of the variables under study. The model is estimated through the maximization of the joint likelihood function (complete formulation), using a non-linear unconstraint optimization algorithm (Nelder-Mead).

The estimation process was tested through a Monte-Carlo experiment. The impulse-response function based on estimated parameters was compared to the values obtained through the use of the original ones.

Regarding the intra-day estimated pattern, some facts brought-up by microstructure theory could be observed:

- Highest intensity (lower durations) were observed close to the opening and closing of trading days;
- For lower duration values: bid-ask spread and instantaneous volatility increase;
- Highest volumes are observed next to the opening of transaction days – what can be explained by the fact that new information accumulated after trading regular time wouldn’t be included in asset prices.

In relation to the adoption of constrains on the complete formulation (different structures), a Likelihood Ratio Test was carried-out. The results point to the acceptance of the variation-free formulation, as suggested by Manganelli. Here, the hypothesis of no causality among trading variables is strongly rejected.

Generally, the new model was successful when dealing with the linear dependence in data. On the other hand, it was observed an excess of dispersion in data, probably due to the nonlinearities – first identified by Engle and Russel (1998), what was not captured.

8 References

Biggs, M.C. "Constrained Minimization Using Recursive Quadratic Programming," Towards Global Optimization (L.C.W.Dixon and G.P.Szergo, eds.), North-Holland, pp.341-349, 1975.

Cox, D. R. “Some Statistical Models Connected with Series of Events (with Discussion)”, Journal of the Royal Statistical Society, Series B, 17, p. 129-164, 1955.

Engle, R. F. and Russell, J. R. “Autoregressive conditional duration: a new model for irregularly spaced transaction data” Econometrica, 1998.

Fernandes, M. and Gramming, J. “A family of autoregressive conditional durantion models” CORE (Center for Operations Research and Econometrics), 2001.

Gaver, D. P. and Lewis, P. A. W., “First-Order Autoregressive Gamma Sequences and Point Processes”, Advances in Applied Probability. 12, 727–745, 1980.

Han, S.P. "A Globally Convergent Method for Nonlinear Programming," J. Optimization Theory and Applications, Vol. 22, p. 297, 1977.

Manganelli, S. “Duration, volume and volatility impact of trades” European Central Bank (Working Paper Series, 2002.

Nelder, J.A. and R. Mead, "A Simplex Method for Function Minimization," Computer J., Vol .7, pp. 308-313, 1965.

Powell, M.J.D. "A Fast Algorithm for Nonlinearly Constrained Optimization Calculations," Numerical Analysis, G.A.Watson ed., Lecture Notes in Mathematics, Springer Verlag, Vol. 630, 1978.

Spendley, W., Hext, G. R. and Himsworth, F. R. “Sequential Application of Simplex Designs in Optimization and Evolutionary Operation”, Technometrics, Vol. 4, p. 441, 1962.

Wold, H. “On Stationary Point Process and Markov Changes”, Skandinavisk Aktuarietidskrist, 31, p. 229-240, 1948.

Zhang, S. Y., Russell J. R. and Tsay, R. S. “A nonlinear autoregressive conditional duration model with applications to financial transaction data” Journal of Econometrics, 2001.

Appendix I

- **Table I (model complete):**

	Parameters Value	Variance	Confidence Interval	H0
a0	0,5655	0,0003	0,0345	Reject
b0	0,6819	0,0007	0,0527	Reject
c0	-0,1939	0,0011	0,0664	Reject
d0	0,2174	0,0022	0,0926	Reject
a1,1	0,4556	0,0005	0,0442	Reject
a2,1	-0,1212	0,0028	0,1036	Reject
a3,1	-0,0619	0,0005	0,0456	Reject
a4,1	-0,2603	0,0002	0,0252	Reject
b1,1	0,5368	0,0028	0,1032	Reject
b2,1	-0,4027	0,0047	0,1345	Reject
b3,1	-0,2179	0,0002	0,0305	Reject
b4,1	0,0160	0,0006	0,0494	Accept
c1,1	-0,0224	0,0002	0,0292	Accept
c2,1	-0,0226	0,0005	0,0417	Accept
c3,1	-0,2159	0,0026	0,1006	Reject
c4,1	-0,0475	0,0018	0,0833	Accept
d1,1	-0,0388	0,0002	0,0278	Reject
d2,1	0,2113	0,0005	0,0449	Reject
d3,1	-0,5421	0,0006	0,0460	Reject
d4,1	-0,2894	0,0044	0,1294	Reject
a1,2	0,0950	0,0031	0,1088	Accept
a2,2	-0,1476	0,0022	0,0912	Reject
a3,2	0,1206	0,0009	0,0583	Reject
a4,2	-0,1085	0,0019	0,0849	Reject
b1,2	0,0416	0,0025	0,0971	Accept
b2,2	-0,0247	0,0009	0,0586	Accept
b3,2	0,2977	0,0000	0,0108	Reject
b4,2	-0,0385	0,0002	0,0284	Reject
c1,2	0,1473	0,0013	0,0707	Reject
c2,2	0,2955	0,0003	0,0364	Reject
c3,2	0,0906	0,0009	0,0596	Reject
c4,2	0,0606	0,0022	0,0926	Accept
d1,2	-0,0003	0,0005	0,0437	Accept
d2,2	-0,1765	0,0010	0,0619	Reject
d3,2	-0,1012	0,0005	0,0427	Reject
d4,2	0,4096	0,0020	0,0870	Reject
b5	0,0614	0,0001	0,0231	Reject
c5	-0,0599	0,0002	0,0256	Reject
c6	0,0231	0,0002	0,0291	Accept
d5	-0,0070	0,0001	0,0206	Accept
d6	0,0168	0,0003	0,0323	Accept
d7	-0,0810	0,0007	0,0529	Reject

	Parameters Value	Variance	Confidence Interval	H0
a5,1	0,1075	0,0002	0,0259	Reject
a6,1	0,2607	0,0004	0,0384	Reject
a7,1	0,0289	0,0011	0,0645	Accept
a8,1	-0,1543	0,0008	0,0560	Reject
b6,1	0,0248	0,0002	0,0279	Accept
b7,1	-0,0867	0,0003	0,0312	Reject
b8,1	-0,0156	0,0002	0,0293	Accept
b9,1	-0,0067	0,0001	0,0140	Accept
c7,1	-0,0847	0,0003	0,0322	Reject
c8,1	0,0192	0,0001	0,0221	Accept
c9,1	0,1554	0,0003	0,0326	Reject
c10,1	0,0578	0,0004	0,0381	Reject
d8,1	-0,1160	0,0005	0,0432	Reject
d9,1	0,0090	0,0002	0,0301	Accept
d10,1	-0,0283	0,0006	0,0498	Accept
d11,1	0,3570	0,0006	0,0486	Reject
a5,2	0,0167	0,0003	0,0320	Accept
a6,2	-0,1942	0,0001	0,0231	Reject
a7,2	-0,1377	0,0001	0,0157	Reject
a8,2	0,1841	0,0001	0,0165	Reject
b6,2	-0,0450	0,0001	0,0186	Reject
b7,2	-0,1424	0,0006	0,0487	Reject
b8,2	-0,0661	0,0003	0,0324	Reject
b9,2	0,0825	0,0004	0,0397	Reject
c7,2	-0,0241	0,0001	0,0221	Reject
c8,2	0,0225	0,0002	0,0282	Accept
c9,2	0,0166	0,0002	0,0281	Accept
c10,2	-0,0581	0,0012	0,0672	Accept
d8,2	-0,1257	0,0002	0,0256	Reject
d9,2	0,0727	0,0001	0,0236	Reject
d10,2	0,0597	0,0005	0,0458	Reject
d11,2	0,2001	0,0013	0,0708	Reject

- **Table II (variation-free):**

	Parameters Value	Variance	Confidence Interval	H0
a_0	0,2568	0,0003	0,0354	Reject
b_0	0,7288	0,0012	0,0690	Reject
c0	0,0641	0,0013	0,0702	Accept
d0	0,2561	0,0003	0,0342	Reject
a1,1	0,3761	0,0019	0,0853	Reject
b2,1	-0,2551	0,0009	0,0576	Reject
c3,1	-0,2793	0,0026	0,0998	Reject
d4,1	-0,4078	0,0012	0,0690	Reject
a1,2	0,2401	0,0013	0,0716	Reject
b2,2	0,2197	0,0028	0,1038	Reject
c3,2	-0,5906	0,0021	0,0903	Reject
d4,2	0,0764	0,0036	0,1170	Accept
b5	0,0397	0,0002	0,0276	Reject
c5	-0,0436	0,0002	0,0268	Reject
c6	-0,0032	0,0002	0,0290	Accept
d5	-0,0313	0,0004	0,0398	Accept
d6	-0,0295	0,0004	0,0411	Accept
d7	-0,1295	0,0006	0,0497	Reject
a5,1	0,1206	0,0001	0,0212	Reject
a6,1	0,1870	0,0001	0,0217	Reject
a7,1	0,1260	0,0008	0,0565	Reject
a8,1	-0,1188	0,0007	0,0507	Reject
b6,1	0,0177	0,0002	0,0288	Accept
b7,1	-0,1063	0,0002	0,0242	Reject
b8,1	0,0429	0,0004	0,0387	Reject
b9,1	-0,0073	0,0008	0,0548	Accept
c7,1	-0,0289	0,0001	0,0235	Reject
c8,1	-0,0188	0,0002	0,0242	Accept
c9,1	0,1755	0,0006	0,0496	Reject
c10,1	-0,1054	0,0007	0,0524	Reject
d8,1	-0,0813	0,0003	0,0359	Reject
d9,1	-0,0195	0,0006	0,0470	Accept
d10,1	-0,1045	0,0007	0,0536	Reject
d11,1	0,5287	0,0010	0,0614	Reject

	Parameters Value	Variance	Confidence Interval	H0
a5,2	0,0583	0,0002	0,0273	Reject
a6,2	-0,1367	0,0002	0,0280	Reject
a7,2	-0,1792	0,0011	0,0649	Reject
a8,2	0,0264	0,0011	0,0652	Accept
b6,2	-0,0152	0,0002	0,0265	Accept
b7,2	-0,0008	0,0002	0,0266	Accept
b8,2	-0,1414	0,0009	0,0589	Reject
b9,2	-0,1902	0,0010	0,0614	Reject
c7,2	-0,0369	0,0001	0,0225	Reject
c8,2	-0,0149	0,0002	0,0275	Accept
c9,2	0,0695	0,0007	0,0514	Reject
c10,2	-0,1009	0,0009	0,0581	Reject
d8,2	-0,1249	0,0004	0,0367	Reject
d9,2	0,0594	0,0004	0,0413	Reject
d10,2	-0,0325	0,0012	0,0689	Accept
d11,2	0,2708	0,0032	0,1113	Reject

- **Table III (individual):**

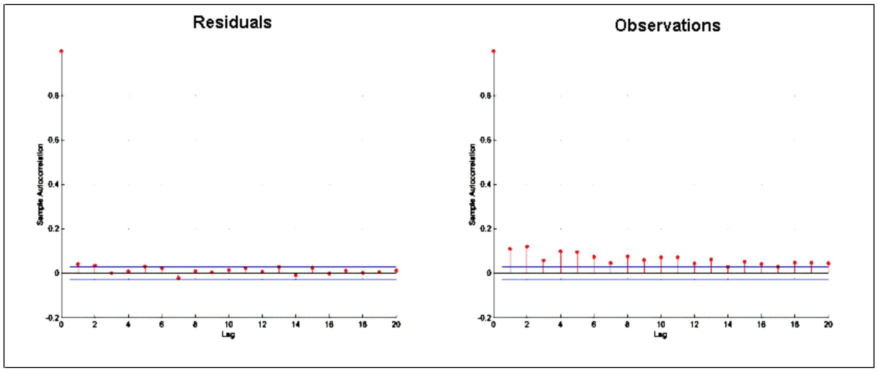
	Parameters Value	Variance	Confidence Interval	H0
a0	0,2250	0,0017	0,0820	Reject
b0	0,3067	0,0008	0,0552	Reject
c0	0,1826	0,0009	0,0595	Reject
d0	0,5954	0,0041	0,1260	Reject
a1,1	0,0791	0,0035	0,1167	Accept
b2,1	0,4615	0,0002	0,0281	Reject
c3,1	0,0756	0,0045	0,1314	Accept
d4,1	-0,8327	0,0074	0,1691	Reject
a1,2	0,5833	0,0031	0,1085	Reject
b2,2	0,1180	0,0010	0,0618	Reject
c3,2	-0,2601	0,0016	0,0772	Reject
d4,2	-0,6936	0,0010	0,0623	Reject
a5,1	0,0901	0,0002	0,0277	Reject
b7,1	-0,1003	0,0002	0,0300	Reject
c9,1	0,3680	0,0015	0,0768	Reject
d11,1	0,4433	0,0013	0,0711	Reject
a5,2	0,1026	0,0003	0,0344	Reject
b7,2	0,0808	0,0002	0,0272	Reject
c9,2	-0,2262	0,0013	0,0703	Reject
d11,2	0,4021	0,0020	0,0886	Reject

Appendix II

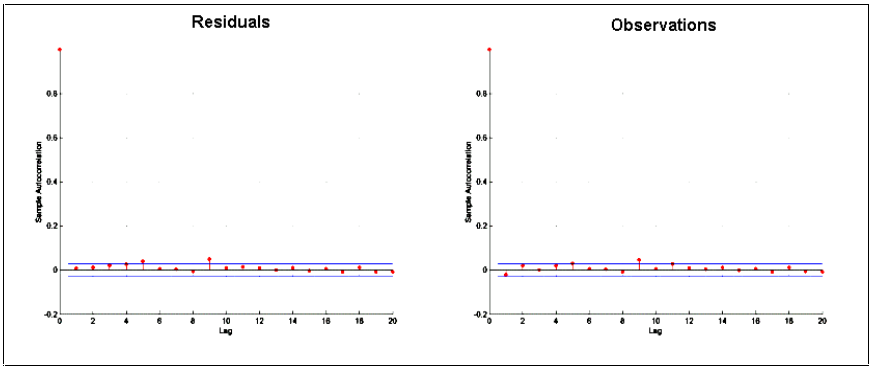
- **Variation-free:**

- **ACF**

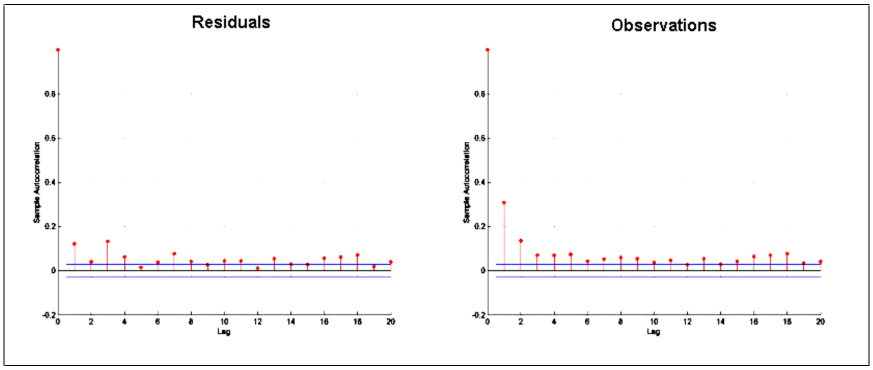
  - Duration



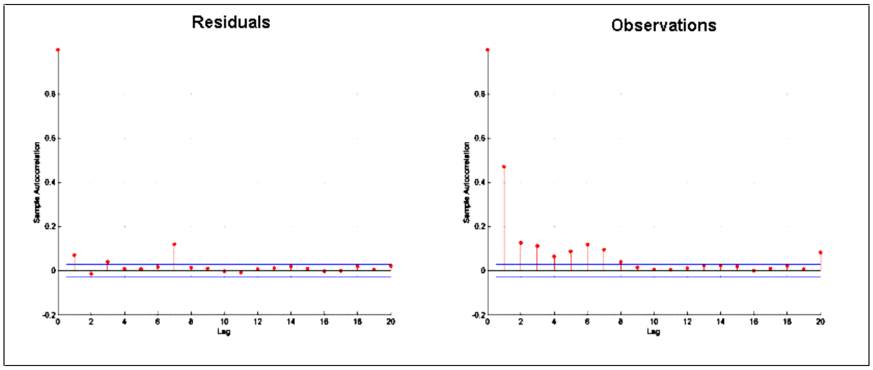
  - Volume



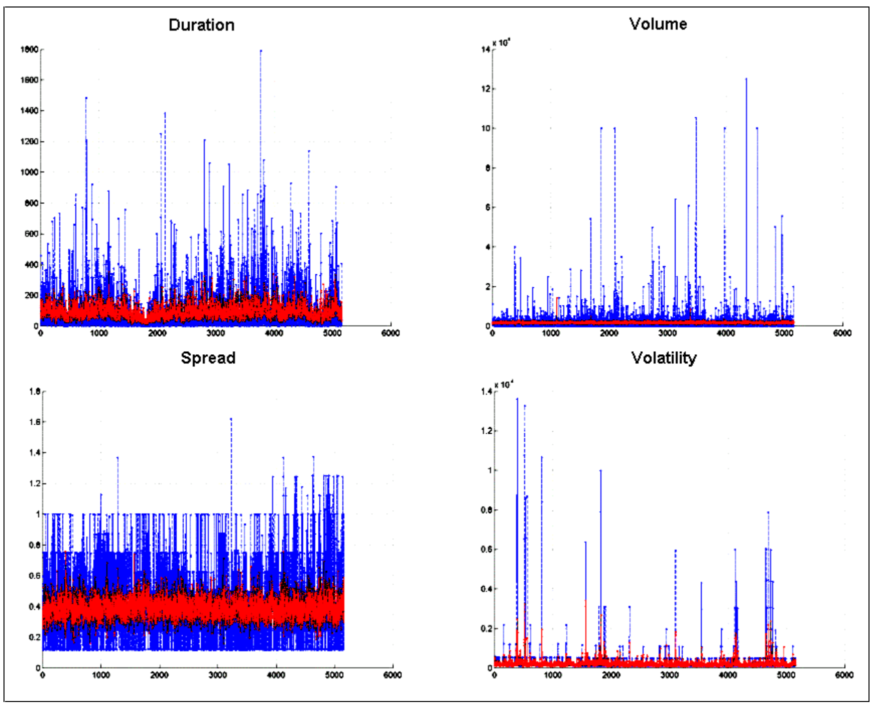
  - Spread



- Volatility



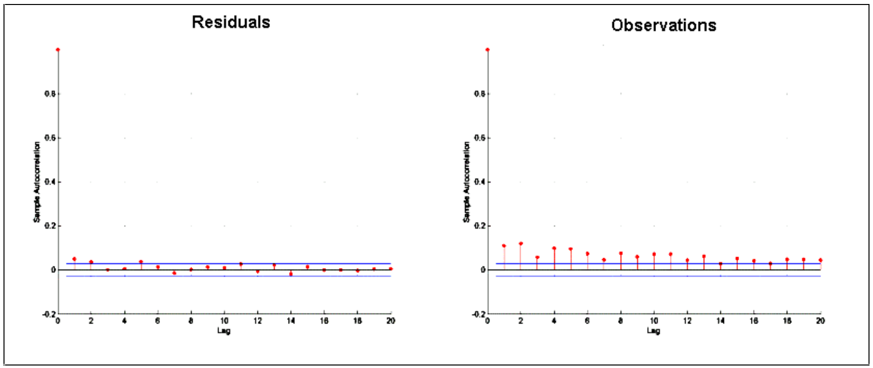
- **Real x Forecasted**



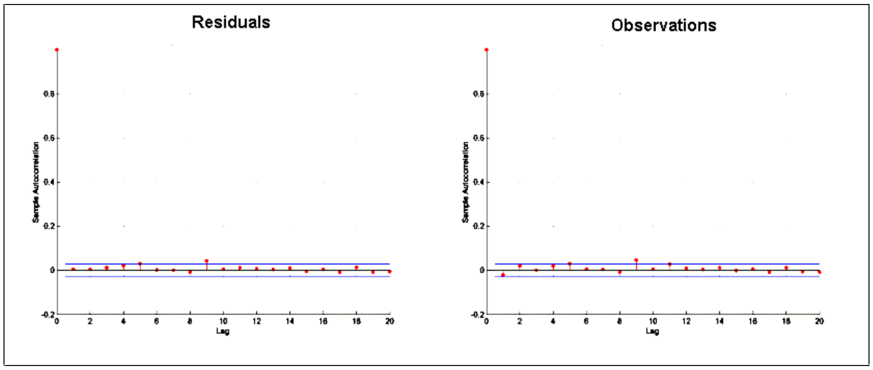
- **Individual:**

- **ACF**

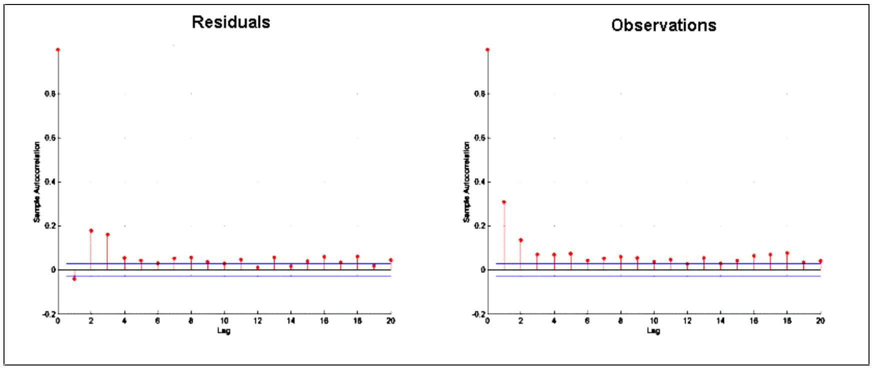
  - Duration



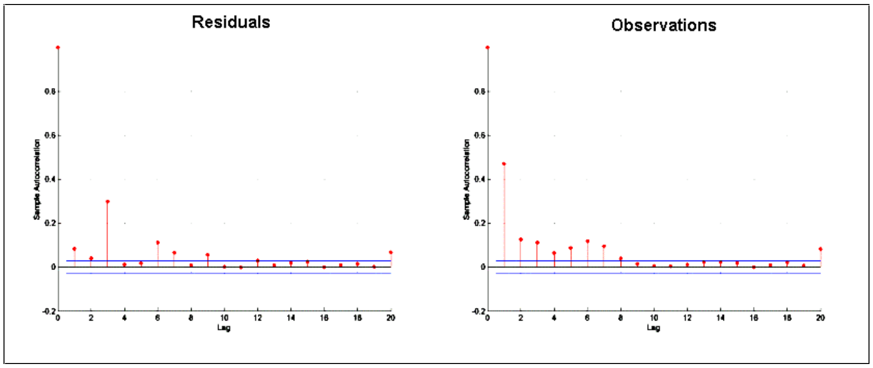
  - Volume



  - Spread



- Volatility



- Real x Forecasted