5
Modelo de Dados de Alta Freqüência

5.1
Introdução

Conforme destacado anteriormente, podem-se observar três fases distintas na modelagem das variáveis relacionadas aos negócios envolvendo instrumentos financeiros. A primeira corresponde aos desenvolvimentos iniciais. A segunda toma por base os conceitos embutidos nos modelos ARCH/GARCH, aplicados para outras variáveis observadas durante as transações financeiras que não a dispersão das variações de preços (volatilidade). Dentre as variáveis mais comumente estudadas, merece destaque o intervalo de tempo entre as transações financeiras – duração.

Ultimamente, a preocupação está relacionada não somente à dinâmica apresentada pelas variáveis de interesse, como também ao grau de influência existente entre elas. A principal meta passa a ser o entendimento das relações verificadas entre as variáveis que caracterizam o processo de conclusão dos negócios realizados no mercado financeiro.

O principal marco desta etapa foi o desenvolvimento de Manganelli (2002), que propôs a modelagem conjunta de diferentes variáveis de interesse (duração, volume e volatilidade) através do uso de modelos vetoriais autoregressivos. Este trabalho foi uma das primeiras tentativas de estudo da dinâmica conjunta de dados de alta freqüência, tendo sido o primeiro no qual informações referentes aos volumes transacionados foram consideradas de forma explícita.

Neste capítulo o trabalho de Manganelli é estendido através da inclusão do *spread* de compra e venda num sistema autoregressivo multivariado. Além disto, é estabelecida uma nova forma funcional para a média condicional, que, neste caso, evolui a partir de uma formulação exponencial, evitando, com isso, a

adoção de restrições nos parâmetros quando da maximização da função de verossimilhança – processo de estimação. Esta nova formulação recebe o nome de EMACM (*Exponential Multivariate Autoregressive Conditional Model*).

5.2
O EMACM

Seja x_i a duração da i-ésima transação financeira realizada, onde $x_i = t_i - t_{i-1}$ (diferença de tempo entre as duas últimas negociações observadas) e z_i o vetor de variáveis explicativas, entendidas como as demais informações (caráter cruzado – *cross section*) contidas de maneira implícita ou explícita nos eventos em estudo, então, pode-se definir:

$$(x_i, z_i) \sim f(x_i, z_i \mid \Omega_{i-1}; \theta) \tag{5.2.1}$$

onde $f(x_i, z_i \mid \Omega_{i-1}; \theta)$ corresponde à função de distribuição de probabilidades conjunta de x_i e z_i, Ω_{i-1} é toda informação contida até o evento anterior ao i-ésimo observado e θ o vetor de parâmetros.

Tomando $z_i' = (\upsilon_i, s_i, y_i)$, onde, volume ($\upsilon_i$), *spread* de compra e venda (s_i) – calculados com base nas ofertas dos formadores de mercado – e retorno (y_i) associados ao i-ésimo evento, tem-se:

$$(x_i, z_i) \sim f(x_i, v_i, s_i, y_i \mid \Omega_{i-1}; \theta)$$

Reescrevendo a função de distribuição de probabilidades conjunta a partir do produto das distribuições de probabilidades condicionais, obtém-se:

$$(x_i, z_i) = g(x_i \mid \Omega_{i-1}; \theta_1) h(v_i \mid x_i, \Omega_{i-1}; \theta_2) k(s_i \mid x_i, v_i, \Omega_{i-1}; \theta_3) l(y_i \mid x_i, v_i, s_i, \Omega_{i-1}; \theta_4) \tag{5.2.2}$$

A formulação estabelecida em (5.2.2) se mostra natural, tendo em vista o uso de modelos estratégicos na literatura referente à teoria de microestruturas de mercado. Por exemplo, Kyle (1985) modela o comportamento de

"investidores informados" tomando por base o efeito das ordens de compra e venda no preço, condicionando a análise às atitudes dos investidores "não-informados" e dos formadores de mercado.

No momento em que a informação se torna pública, passa a ocorrer uma grande pressão de oferta/demanda causada, principalmente, pelos formadores de mercado. Uma vez que o intervalo de tempo entre transações financeiras, bem como o volume movimentado nas mesmas, pode indicar a atuação de investidores com informação privilegiada, os formadores de mercado irão considerar este aspecto na definição dos preços a serem ofertados, de modo a prevenir perdas. Assim, os "investidores informados" passarão a dividir as transações financeiras em lotes menores, reduzindo o volume e aumentando o número de transações por unidade de tempo – intensidade. Tal atitude dificulta a identificação das mesmas por parte dos formadores de mercado, postergando possíveis alterações nos *spreads* e, conseqüentemente, no nível de preços praticados.

Com base no que fora estabelecido por Kyle, torna-se clara a opção feita em favor da relação estabelecida pela equação (5.2.2).

Definidas cada uma das componentes do sistema, os modelos podem ser determinados separadamente:

(1) ACD (*Autoregressive Conditional Duration*):

$$x_i = \psi_i \,.\, \varepsilon_i \quad \rightarrow \quad \varepsilon_i \sim \exp(1) \tag{5.2.3}$$

$$\psi_i = E\left(x_i \middle| \, \Omega_i ; \theta_x\right) \tag{5.2.4}$$

- Espaço de definição da variável de interesse: Reais positivos;

(2) ACV (*Autoregressive Conditional Volume*): definido de forma análoga aos modelos de duração, devido às similaridades existentes.

$$v_i = \phi_i \,.\, \eta_i \quad \rightarrow \quad \eta_i \sim \exp(1) \tag{5.2.5}$$

$$\phi_i = E\left(\upsilon_i \middle| \Omega_i; \theta_\upsilon\right) \tag{5.2.6}$$

- Espaço de definição da variável de interesse: Reais positivos;

(3) ACS (*Autoregressive Conditional Spread*): definido de forma análoga aos modelos de duração, devido às similaridades existentes.

$$s_i = P_{venda,\,i} - P_{compra,\,i}$$

Onde, P_{compra} corresponde ao preço ofertado pelos formadores de mercado para a realização de uma operação de compra no instante de tempo referente ao fechamento da i-ésima transação financeira (P_{venda} é o análogo para uma operação de venda).

$$s_i = \varphi_i \,.\, \varpi_i \quad \rightarrow \quad \varpi_i \sim \exp(1) \tag{5.2.7}$$

$$\varphi_i = E\left(s_i \middle| \Omega_i; \theta_s\right) \tag{5.2.8}$$

- Espaço de definição da variável de interesse: Reais positivos;

(4) GARCH: modelo escolhido para analisar a dinâmica da volatilidade intradiária.

$$y_i = \sigma_i \,.\, \varsigma_i \quad \rightarrow \quad \varsigma_i \sim N(0,1) \tag{5.2.9}$$

$$\sigma_i^2 = E\left(y_i^2 \middle| \Omega_i; \theta_y\right) \tag{5.2.10}$$

- Espaço de definição da variável de interesse: Reais positivos;

Assim, para o modelo EMACM (p, q) a formulação final da média condicional será:

$$\ln(\mu_i) = \gamma + \sum_{k=1}^{q} A_k \ln(\mu_{i-k}) + \sum_{m=0}^{p} B_m \ln(\tau_{i-m}) \tag{5.2.11}$$

onde, μ_i' = (ψ_i, ϕ_i, φ_i, σ^2_i), τ_i' = (x_i, υ_i, s_i, y^2_i), γ é um vetor de coeficientes e A_1,..., A_q e B_1,..., B_p são matrizes de coeficientes de cada um dos processos estocásticos que compõem o sistema.

Assim, tem-se a formulação geral do modelo completo (variante):

- **Equação associada às variáveis de interesse**

$$\begin{bmatrix} x_i \\ \upsilon_i \\ s_i \\ y_i \end{bmatrix} = \begin{bmatrix} \psi_i & 0 & 0 & 0 \\ 0 & \phi_i & 0 & 0 \\ 0 & 0 & \varphi_i & 0 \\ 0 & 0 & 0 & \sigma_i \end{bmatrix} \cdot \begin{bmatrix} \varepsilon_i \\ \eta_i \\ \varpi_i \\ \varsigma_i \end{bmatrix} \tag{5.2.12}$$

onde, ε_i, η_i e ω_i ~ exp(1) e ζ_i ~ N(0,1).

- **Equação da média condicional:**

$$\ln\begin{bmatrix} \psi_i \\ \phi_i \\ \varphi_i \\ \sigma^2_i \end{bmatrix} = \begin{bmatrix} a_0 \\ b_0 \\ c_0 \\ d_0 \end{bmatrix} + \sum_{l=1}^{q} \begin{bmatrix} a_1^{(l)} & a_2^{(l)} & a_3^{(l)} & a_4^{(l)} \\ b_1^{(l)} & b_2^{(l)} & b_3^{(l)} & b_4^{(l)} \\ c_1^{(l)} & c_2^{(l)} & c_3^{(l)} & c_4^{(l)} \\ d_1^{(l)} & d_2^{(l)} & d_3^{(l)} & d_4^{(l)} \end{bmatrix} \ln\begin{bmatrix} \psi_{i-l} \\ \phi_{i-l} \\ \varphi_{i-l} \\ \sigma^2_{i-l} \end{bmatrix}$$
$$+ \begin{bmatrix} 0 & 0 & 0 & 0 \\ b_5 & 0 & 0 & 0 \\ c_5 & c_6 & 0 & 0 \\ d_5 & d_6 & d_7 & 0 \end{bmatrix} \ln\begin{bmatrix} x_i \\ \upsilon_i \\ s_i \\ y^2_i \end{bmatrix} + \sum_{m=1}^{p} \begin{bmatrix} a_5^{(m)} & a_6^{(m)} & a_7^{(m)} & a_8^{(m)} \\ b_6^{(m)} & b_7^{(m)} & b_8^{(m)} & b_9^{(m)} \\ c_7^{(m)} & c_8^{(m)} & c_9^{(m)} & c_{10}^{(m)} \\ d_{11}^{(m)} & d_{12}^{(m)} & d_{13}^{(m)} & d_{14}^{(m)} \end{bmatrix} \ln\begin{bmatrix} x_{i-m} \\ \upsilon_{i-m} \\ s_{i-m} \\ y^2_{i-m} \end{bmatrix} \tag{5.2.13}$$

O fato de a matriz B_0 ser diagonal inferior, com elementos nulos na diagonal principal, confere ao sistema uma dinâmica interativa que estabelece uma ordem de precedência na atualização das componentes.

Com base na equação (5.2.13), pode-se inferir sobre a estrutura do sistema mediante a adoção de diferentes restrições. Três estruturas são sugeridas:

- **Modelo completo**: a estrutura dos coeficientes das matrizes é exatamente como apresentado na equação (5.2.13). Neste caso, a média condicional e as variáveis contemporâneas e defasadas exercem influência sobre a dinâmica do sistema como um todo. A relação de causalidade é explicitamente modelada.
- **"Livre de variação" (*variation-free*)**: as matrizes A_1,..., A_q são diagonais. As médias condicionais das componentes do sistema não exercem influência entre si.
- **Individual**: a matriz B_0 é nula e as demais presentes em (5.2.13) são diagonais. Nesta, as formulações individuais são obtidas (ACD, ACV, ACS e GARCH).

Com base nas equações (5.2.2), (5.2.3), (5.2.5), (5.2.7) e (5.2.9), pode-se escrever a função de verossimilhança:

$$L\left(x, \upsilon, s, y \middle| I_N\right) = \prod_{i=1}^{N} \left[\begin{array}{l} \frac{1}{\psi_i} \cdot \exp\left(-\frac{x_i}{\psi_i}\right) \cdot \frac{1}{\phi_i} \cdot \exp\left(-\frac{\upsilon_i}{\phi_i}\right) \cdot \frac{1}{\varphi_i} \cdot \exp\left(-\frac{s_i}{\varphi_i}\right) \\ \cdot \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{y_i^2}{2\sigma_i^2}\right) \end{array} \right] \tag{5.2.14}$$

e log-verossimilhança:

$$l\left(x, \upsilon, s, y \middle| I_N\right) = \sum_{i=1}^{N} \left\{ \begin{array}{l} \ln\left(\frac{1}{\psi_i}\right) - \left(\frac{x_i}{\psi_i}\right) + \ln\left(\frac{1}{\phi_i}\right) - \left(\frac{\upsilon_i}{\phi_i}\right) + \ln\left(\frac{1}{\varphi_i}\right) - \left(\frac{s_i}{\varphi_i}\right) \\ -\frac{1}{2} \cdot \left[\ln(2\pi) + \ln\left(\sigma_i^2\right) + \left(\frac{y_i^2}{\sigma_i^2}\right) \right] \end{array} \right\} \tag{5.2.15}$$

Onde, as médias condicionais das variáveis de interesse (ψ_i, ϕ_i, φ_i e σ^2_i) são definidas por (5.2.13).

5.3
Ajuste sazonal (padrão intradiário)

Muitas das variáveis analisadas podem apresentar um padrão sazonal ao longo do dia de negociação – padrão sazonal intradiário. Conforme proposto por Engle (1998), todo comportamento periódico ou cíclico deve ser removido antes da estimação do modelo, de modo a evitar o aparecimento de correlações espúrias. Assim, pode-se definir as seguintes séries "sazonalmente ajustadas" das nossas variáveis de interesse.

$$x_i^* = x_i \Big/ \lambda(x_i, t_i) \quad v_i^* = v_i \Big/ \lambda(\upsilon_i, t_i) \quad s_i^* = s_i \Big/ \lambda(s_i, t_i) \quad y_i^{2*} = y_i^2 \Big/ \lambda(y_i^2, t_i) \tag{5.3.1}$$

Onde, x_i^*, υ_i^*, s_i^*, y_i^{2*} são, respectivamente, as variáveis referentes à duração, volume, *spread* e retorno. Nestas, o comportamento cíclico foi removido a partir do emprego de um termo multiplicativo (ajuste sazonal).

No presente trabalho, o ajuste sazonal se dá mediante a estimação "*off-line*" de uma *spline* cúbica natural. Para tal, o horizonte de tempo entre a abertura do pregão e seu fechamento foi dividido em intervalos regulares de uma hora de duração. Visando proporcionar maior flexibilidade, acrescentou-se um nó extra ao final do pregão.

Assim, a função responsável pelo ajuste sazonal nas componentes do modelo será:

$$\lambda(t_{i-1}) = \sum_{j=1}^{K} I_j \left[c_j + d_{1,j}(t_{i-1} - k_{j-1}) + d_{2,j}(t_{i-1} - k_{j-1})^2 + d_{3,j}(t_{i-1} - k_{j-1})^3 \right] \tag{5.3.2}$$

Onde,

K – número de segmentos;
I – variável indicadora do j-ésimo segmento da *spline* (I_j =1 se $k_{j-1} < t_{i-j} < k_j$ e I_j = 0, caso contrário).

5.4
Estimação dos modelos

Com relação ao processo de estimação dos modelos, foram utilizados dois algoritmos distintos. São eles:

- Programação seqüencial quadrática: estimação dos coeficientes da função que caracteriza o padrão sazonal intradiário e do modelo linear individual (análogo ao proposto por Engle (1998) para duração).
- Método simplex de Nelder-Mead: estimação dos coeficientes do modelo conjunto (sistema de equações), em razão das descontinuidades da função de log-verossimilhança.

5.4.1
Programação seqüencial quadrática

Desenvolvido com base nos trabalhos de Biggs (1975), Han (1977) e Powell (1978), o método pode ser entendido em linhas gerais como uma *proxy* do método de Newton (programação não-linear sem restrições), para problemas com restrições.

Algoritmo: A cada interação, o Hessiano da função é calculado, ou melhor, atualizado utilizando-se o método BFGS (*quasi-Newton*). O Hessiano é então utilizado na formulação de um “sub-problema” de programação quadrática (“linearização” das restrições não-lineares), cuja solução é tomada como referência (direção) para o procedimento de busca linear subseqüente, dando início à nova interação.

5.4.2
Método Simplex de Nelder-Mead

Introduzido por Nelder e Mead (1965), a principal idéia do método é a determinação do ponto de mínimo de uma função de N variáveis, a partir da utilização de um simplex de N+1 vértices, o qual varia dinamicamente, de acordo com determinadas regras estratégicas – Spendley (1962).

Neste método, o simplex possui um mecanismo auto-adaptativo, alongando diante de planos inclinados, mudando de direção ao encontrar um "vale" e contraindo na vizinhança de um ponto de mínimo.

Cabe ressaltar que o critério de parada, inicialmente proposto por Nelder e Mead, levava em consideração a aplicabilidade do método em problemas estatísticos envolvendo a maximização da função de verossimilhança, nos quais os parâmetros desconhecidos são expressos no modelo por meio de relações não-lineares.

Algoritmo: Seja o problema de minimização de uma determinada função de N variáveis, sem restrições. Tome P_0, P_1,..., P_N os N+1 pontos no espaço N-dimensional que definem o simplex corrente. Define-se y_i como sendo o valor da função em P_i, $y_h = \max(y_i)$ para i = 0, ..., N e $y_l = \min(y_i)$ para i = 0, ..., N.

Adicionalmente, seja P_{hat} o centróide da região formada pelos P_i's, onde i é diferente de h e $[P_i\ P_j]$ a distância de P_i até P_j. Para cada estágio do processo P_h é substituído por um novo ponto; três operações são utilizadas – reflexão, contração e expansão.

A reflexão de P_h é representada por P* e suas coordenadas são definidas pela relação:

$$P^* = (1+\alpha).P_{hat} - \alpha . P_h \qquad (5.4.1)$$

onde, α é uma constante positiva (coeficiente de reflexão).

Desse modo, o ponto P* estará situado na linha que liga os pontos P_h e P_{hat} e [P* P_{hat}] = a . [P_h P_{hat}]. Se $y_l < y^* < y_h$, então P_h será substituído por P* e um novo simplex é gerado.

Caso $y^* < y_l$, ou seja, caso a reflexão tenha ocasionado um novo ponto de mínimo, P* é expandido para P** de acordo com a seguinte relação:

$$P^{**} = \gamma \,.\, P^{*} + (1-\gamma)\,.\, P_{hat} \tag{5.4.2}$$

O coeficiente de expansão γ (valor maior do que 1), corresponde a razão entre as distâncias [P** P_{hat}] e [P* P_{hat}]. Se $y^{**} < y_l$, P_h é substituído por P** e o processo é reiniciado. Porém, se $y^{**} > y_l$, então a expansão falhou e, antes de reiniciar o processo, P_h deve ser substituído por P*.

Caso, ao refletir P para determinação de P*, $y^* > y_i$, para todo i diferente de h, então se deve definir um novo P_h como sendo ou o último valor de P_h ou P* (aquele que corresponder a um menor valor de y) e calcular:

$$P^{**} = \beta \,.\, P_h + (1-\beta)\,.\, P_{hat} \tag{5.4.3}$$

O coeficiente de contração b (valor entre 0 e 1), corresponde à razão entre [P** P_{hat}] e [P P_{hat}]. Desse modo, P_h será substituído por P**, a não ser que $y^{**} > \min(y_h, y^{**})$. Caso tal fato ocorra, os pontos P_i's são substituídos por (P_i + P_i) / 2 e o processo é reiniciado.

O processo termina quando o valor do diâmetro formado pelos pontos que compõem o simplex for inferior a um determinado valor pré-estabelecido. Esta idéia está associada à medida estatística de dispersão e o sucesso em utilizar tal critério está associado ao fato de a região do simplex não se tornar tão pequena com relação à curvatura da superfície, até que o ponto de mínimo "final" seja atingido.

Obtenção do Hessiano: A utilização do método de Nelder-Mead faz com que na resolução do problema de otimização não se faça necessária a determinação do Hessiano.

Nas aplicações em estatística, o Hessiano é de vital importância, uma vez que a determinação da Matriz de Informação de Fisher se dá a partir deste, possibilitando assim a realização de testes de hipóteses. De modo a contornar este problema, Nelder e Mead propõem a utilização do método de Spendley (1962), o qual ajusta uma superfície quadrática na região formada pelos N+1 pontos que formam o simplex na vizinhança da solução ótima.

5.5
Capacidade de identificação do modelo

Para avaliar a capacidade de identificação do método proposto, foram executados experimentos a partir do resultado de simulações (Simulação de Monte Carlo – SMC) de um dado processo estocástico pré-estabelecido – EMACM (1,1). Foram produzidas 45 diferentes realizações do processo, contendo 1000 observações cada. A estimação dos parâmetros se deu a partir da maximização da função de verossimilhança, conforme apresentado em 5.4. A tabela 1 apresenta os resultados obtidos.

Coeficientes	Valor	$q_{5\%}$	Média	$q_{95\%}$	Coeficientes	Valor	$q_{5\%}$	Média	$q_{95\%}$
a0	0,16	-0,03	0,12	0,26	c5	0,21	0,15	0,20	0,26
b0	-0,19	-0,38	-0,28	-0,18	c6	0,27	0,23	0,28	0,32
c0	-0,20	-0,27	-0,17	-0,08	d5	0,20	0,15	0,20	0,25
d0	-0,27	-0,39	-0,27	-0,16	d6	0,23	0,17	0,23	0,30
a1,1	0,13	0,06	0,13	0,20	d7	0,34	0,26	0,34	0,41
a2,1	-0,11	-0,23	-0,12	-0,01	a5,1	0,26	0,19	0,27	0,36
a3,1	-0,32	-0,83	-0,40	0,02	a6,1	-0,19	-0,26	-0,16	-0,06
a4,1	0,30	0,17	0,31	0,44	a7,1	-0,68	-0,75	-0,69	-0,63
b1,1	0,11	0,06	0,12	0,18	a8,1	-0,13	-0,08	-0,06	-0,04
b2,1	0,17	0,06	0,14	0,23	b6,1	0,09	0,03	0,10	0,18
b3,1	-0,09	-0,33	-0,13	0,06	b7,1	-0,58	-0,65	-0,57	-0,49
b4,1	-0,24	-0,30	-0,23	-0,15	b8,1	0,47	0,41	0,47	0,53
c1,1	-0,07	-0,12	-0,07	-0,01	b9,1	-0,11	-0,07	-0,06	-0,04
c2,1	0,11	0,04	0,10	0,16	c7,1	0,08	0,03	0,08	0,14
c3,1	-0,06	-0,22	-0,10	0,03	c8,1	0,15	0,12	0,17	0,22
c4,1	0,11	0,06	0,12	0,18	c9,1	-0,16	-0,23	-0,16	-0,10
d1,1	0,12	0,06	0,14	0,21	c10,1	0,04	0,00	0,02	0,04
d2,1	0,27	0,20	0,28	0,36	d8,1	0,29	0,18	0,29	0,39
d3,1	0,34	0,03	0,32	0,61	d9,1	0,01	0,00	0,01	0,03
d4,1	0,23	0,14	0,25	0,35	d10,1	-0,39	-0,48	-0,41	-0,35
b5	0,36	0,33	0,36	0,39	d11,1	0,06	0,01	0,03	0,05

Tabela 1 – Experimento de SMC (nível de significância: 90% bi-caudal)

Os processos estimados a partir dos resultados das simulações são comparados com o processo gerador dos dados (*Data Generating Process* - DGP) através da análise da função impulso-resposta do sistema.

As figuras 5.1, 5.2, 5.3 e 5.4 apresentam a comparação dos resultados da função impulso-resposta, obtidos com base nos parâmetros originais, linha azul, e dos estimados a partir dos resultados das simulações, linha vermelha.

Além desta, tomando por base a distribuição empírica dos coeficientes estimados, é conduzido um teste de hipótese multivariado (T^2 de Hotelling) de modo a ratificar a significância estatística da equivalência da média dos estimadores com relação aos valores reais dos coeficientes.

- **T^2 de Hotelling:** $$\left(\bar{x}-\mu\right)^T \Sigma^{-1}\left(\bar{x}-\mu\right)=\frac{(N-1)P}{(N-P)}F_{P,N-P} \qquad (5.5.1)$$

Onde, N é o número de elementos da amostra, P é o número de parâmetros a serem estimados, $\bar{x}$ é a média amostral, μ é a média real e Σ é a matriz de variância-covariância da amostra.

Substituindo os valores obtidos a partir do experimento de simulação, tem-se:

$$557.34 = \frac{(45-1)42}{(45-42)} F_{42,3}(\alpha = 0.05) \Leftrightarrow 557.34 << 5291.44$$

Como o valor calculado se mostra inferior ao limite estabelecido para o nível de significância de 95%, então, a hipótese de equivalência entre as médias é fortemente aceita (hipótese nula ou H_0 → as médias são estatisticamente iguais), fato que sugere a boa aderência do processo de estimação.

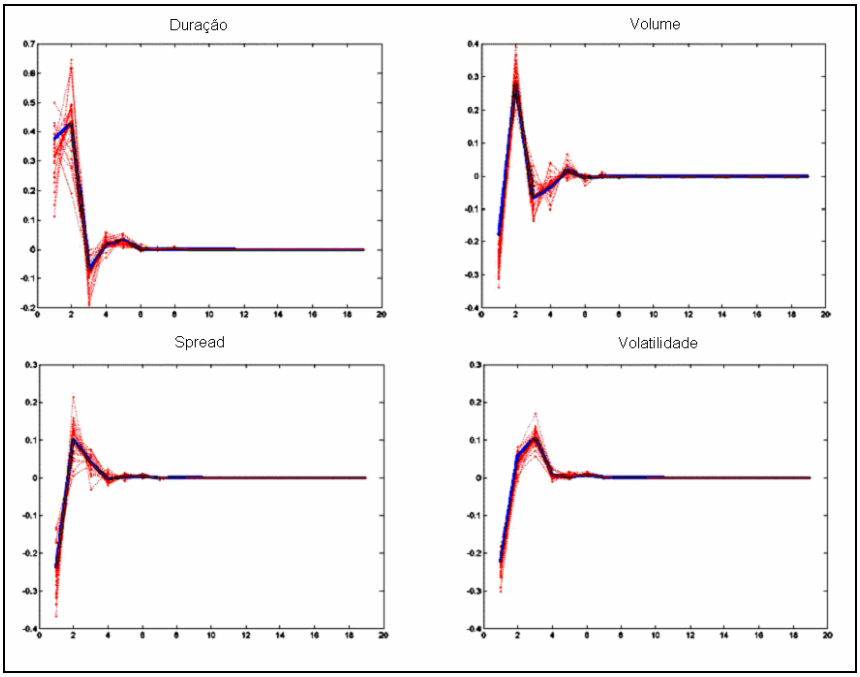

Figura 5.1: Resposta da duração devido a impulso nas componentes

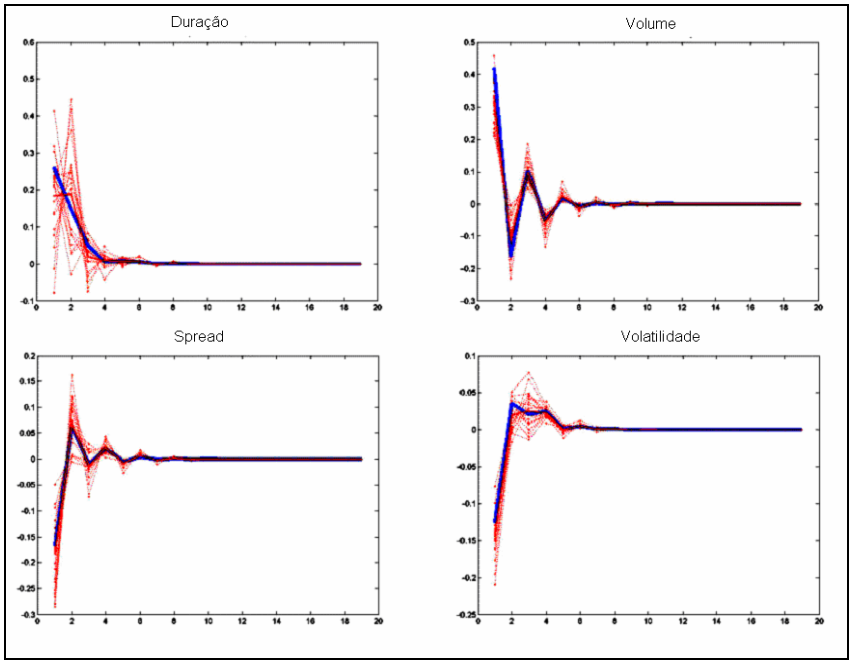

Figura 5.2: Resposta do volume devido a impulso nas componentes

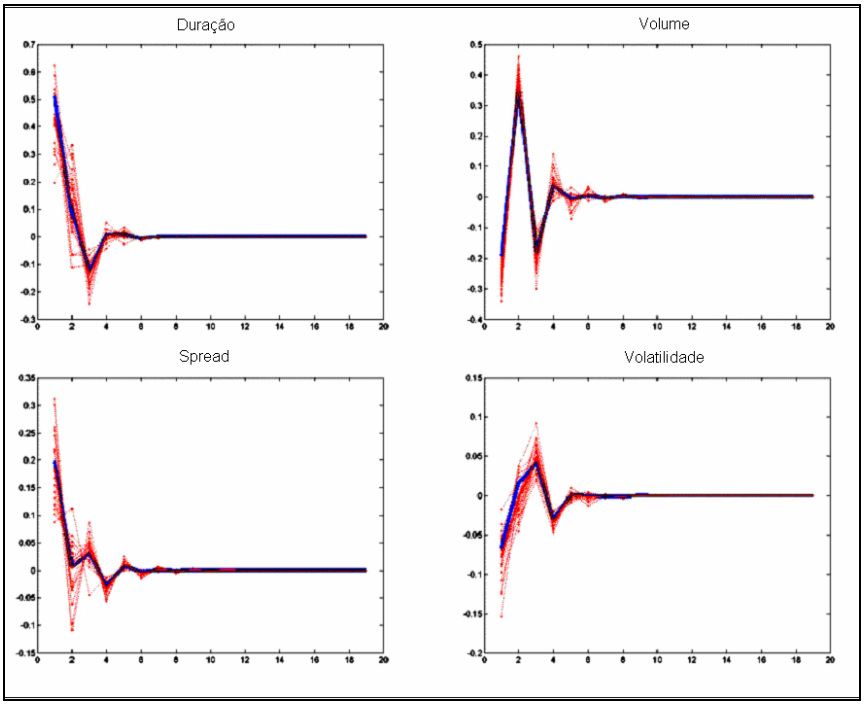

Figura 5.3: Resposta do spread devido a impulso nas componentes

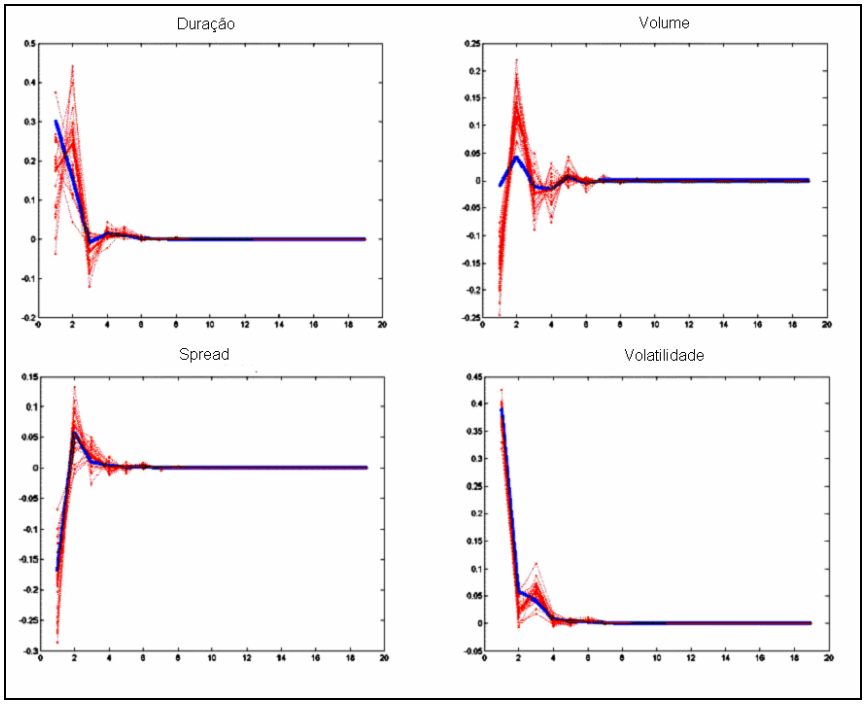

Figura 5.4: Resposta da volatilidade devido a impulso nas componentes

5.6
Análise empírica

5.6.1
Base de dados

A base de dados utilizada na tese foi construída e disponibilizada originalmente por Joel Hasbrouck e NYSE – *Trades, Orders Reports and Quotes* (TORQ). Os dados dizem respeito aos negócios envolvendo ações da IBM, ocorridos entre 1 de Novembro de 1990 e 3 de Dezembro de 1990.

A base de dados contempla todas as informações relevantes ao processo de fechamento das operações de compra e venda (i.e., *bid price*, *ask price*, preço de fechamento, horário do negócio e volume financeiro) realizadas no referido período, durante o horário regular de negócios – 9:30 AM - 4:00 PM (não contempla *after-market*).

Como o presente estudo toma por base informações negócio-a-negócio com variação de preço, torna-se necessária a realização de algumas transformações na base de dados original. Desta forma, tem-se:

- **Duração:**
  - Se o preço do negócio “i” for igual ao preço do negócio “i-1”, então as durações são somadas;
  - Se uma dada operação apresenta duração nula, então o registro da transação é removido da base.
- **Volume:**
  - Se o preço do negócio “i” for igual ao preço do negócio “i-1”, então o volume “i” será dado pela média aritmética dos volumes das duas transações;
- **Spread:**
  - *Spread* referente ao negócio “i” será dado pela diferença entre o *bid price* e o *ask price*;
  - Se o preço do negócio “i” for igual ao preço do negócio “i-1”, então o *spread* “i” será dado pela média ponderada (volume) dos *spreads* “i” e “i-1”.
- **Outras mudanças relevantes e considerações:**
  - 23 de Novembro de 1990: removido devido a uma interrupção de aproximadamente 1 hora e 15 minutos no pregão;
  - Adoção do valor unitário do *tick* como parâmetro básico na determinação das séries com alteração de preço (US$ 0.125);
  - Transações ocorridas nos primeiros vinte minutos de pregão não foram consideradas para fins de estimação (9:30 AM – 9:50 AM), devido a atraso na abertura e aos chamados efeitos dos “primeiros negócios”;

- Para cada dia, o valor inicial da média condicional das variáveis de interesse (sem efeitos sazonais) será dado pela média aritmética das realizações observadas entre 9:50 AM e 10:00 AM. Caso não exista nenhuma observação neste horário, será adotado valor 1 (um).

5.7
Testes empíricos

O conjunto de dados utilizados contabiliza 5806 diferentes transações financeiras. As figuras 1, 2, 3 e 4 do apêndice III apresentam os resultados da análise descritiva das séries de interesse. Os dados amostrais apresentam boa aderência às funções de distribuição de probabilidades de cada uma das componentes do sistema. Além disso, evidenciam-se alguns fatos estilizados já apontados oportunamente (i.e., presença de *clusters* nas séries).

A primeira etapa do experimento corresponde à estimação[2] do modelo EMACM (2,2), conforme descrito em 5.4. Neste as três estruturas anteriormente enumeradas são consideradas (tabelas 5, 6 e 7 do apêndice II trazem os resultados).

Como mencionado, antes de iniciar o processo de estimação é necessário remover a componente periódica (padrão sazonal intradiário). Dessa forma, considerando o processo definido em 5.3, a estimativa da função responsável pelo comportamento cíclico é obtida. A figura 5.5 apresenta os principais resultados para cada uma das variáveis consideradas.

[2] A determinação do Hessiano é realizada com base nos estudos de Spendley et al (1962).

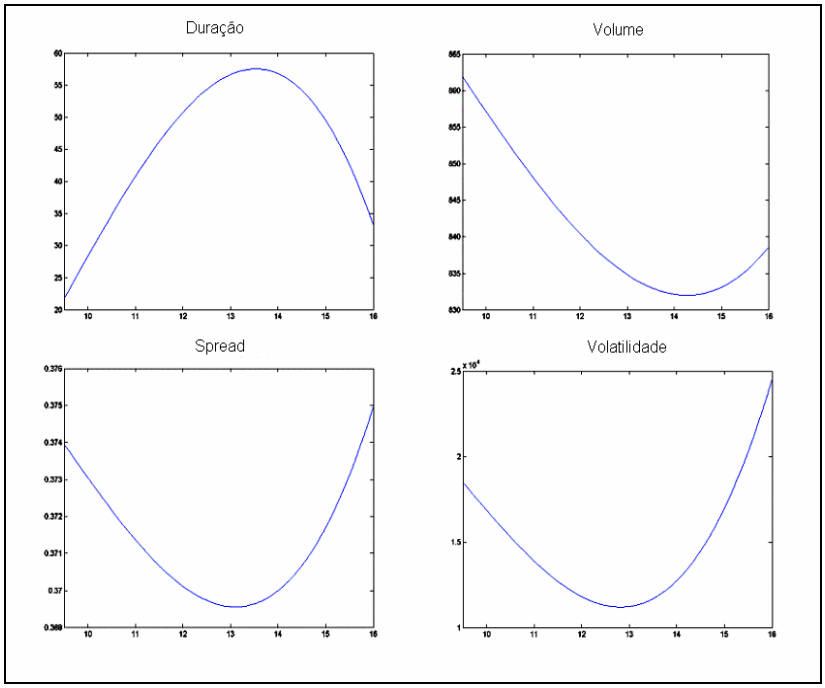

Figura 5.5: Padrão sazonal intradiário das componentes do sistema

Conforme destacado por Engle e Russell (1998), as transações financeiras ocorrem com maior intensidade (baixa duração) nos instantes de tempo próximos da abertura e fechamento do pregão. Adicionalmente, pode-se observar que tanto a diferença de preço entre as ofertas de compra e venda executadas pelos formadores de mercado, quanto a volatilidade dos preços, aumentam em conseqüência deste fato.

Com relação ao padrão sazonal referente ao volume das negociações, convém destacar que os maiores valores observados ocorrem próximos da abertura do pregão, refletindo assim o fato de que as novas informações (efeito "*after-market*") ainda não foram incorporadas aos preços dos ativos.

Após remover os efeitos sazonais, o sistema pode ser estimado através do uso do método Simplex de Nelder e Mead.

5.7.1
Principais resultados

Os principais resultados da análise do processo de estimação para cada uma das três formulações propostas são apresentados a seguir.

- Modelo completo:
  - ACF:
    - Duração: o modelo captura a dependência linear observada nos dados.

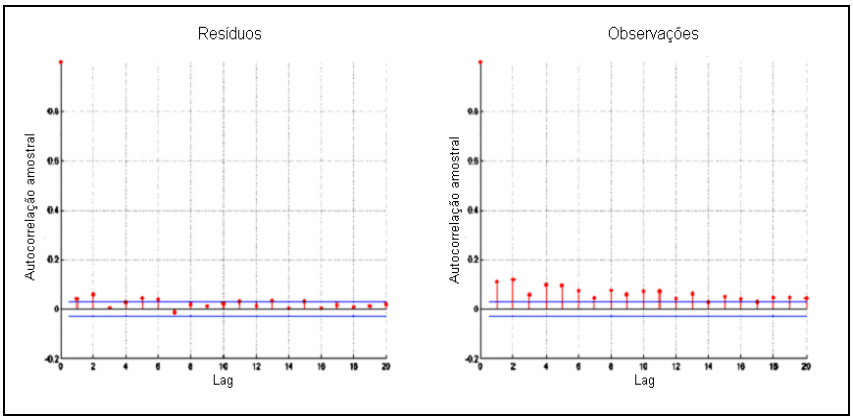

Figura 5.6: ACF duração (resíduos x observações)

- Volume: tanto os dados originais quanto os resíduos não apresentam dependência linear. Entretanto, a hipótese dos parâmetros não serem estatisticamente significantes é fortemente rejeitada (verossimilhança conjunta).

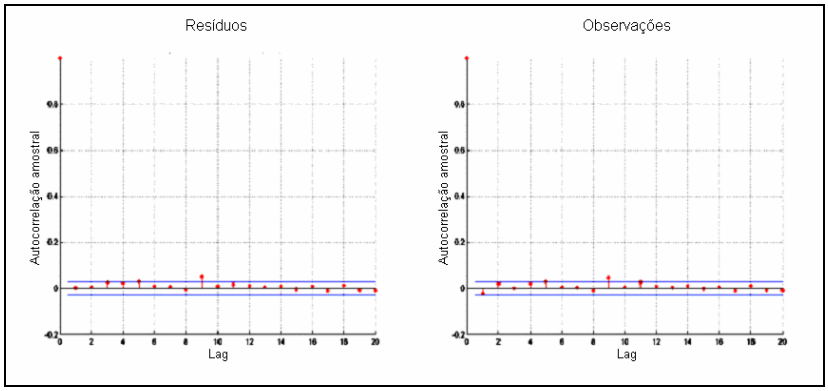

Figura 5.7: ACF volume (resíduos x observações)

- *Spread* de compra e venda: o modelo reduz a dependência linear observada.

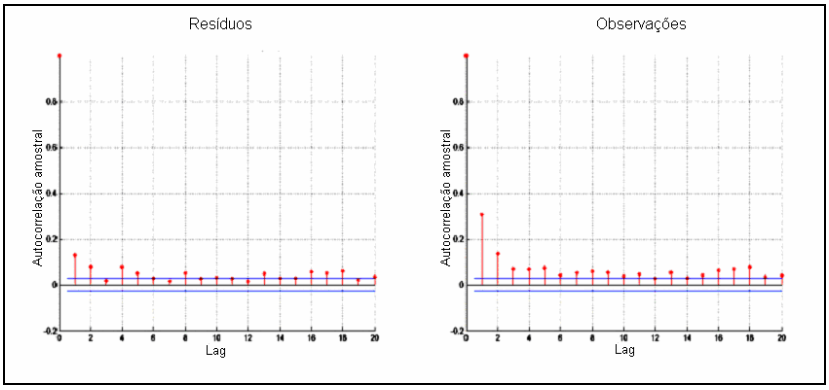

Figura 5.8: ACF *spread* (resíduos x observações)

- Volatilidade: pode-se observar uma forte dependência linear de primeira ordem na volatilidade instantânea. O modelo reduz a intensidade desta dependência, mas não a elimina totalmente.

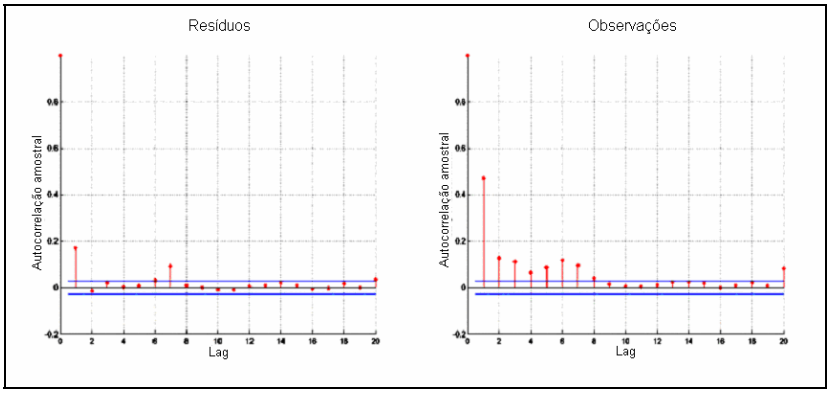

Figura 5.9: ACF volatilidade instantânea (resíduos x observações)

o Previsto x realizado: a figura 5.10 apresenta os gráficos que comparam as previsões um passo à frente das variáveis de interesse com os valores observados das mesmas. Pode-se observar excesso de dispersão nos resíduos, fato não capturado pelo modelo. Isto se deve principalmente às não-linearidades existentes.

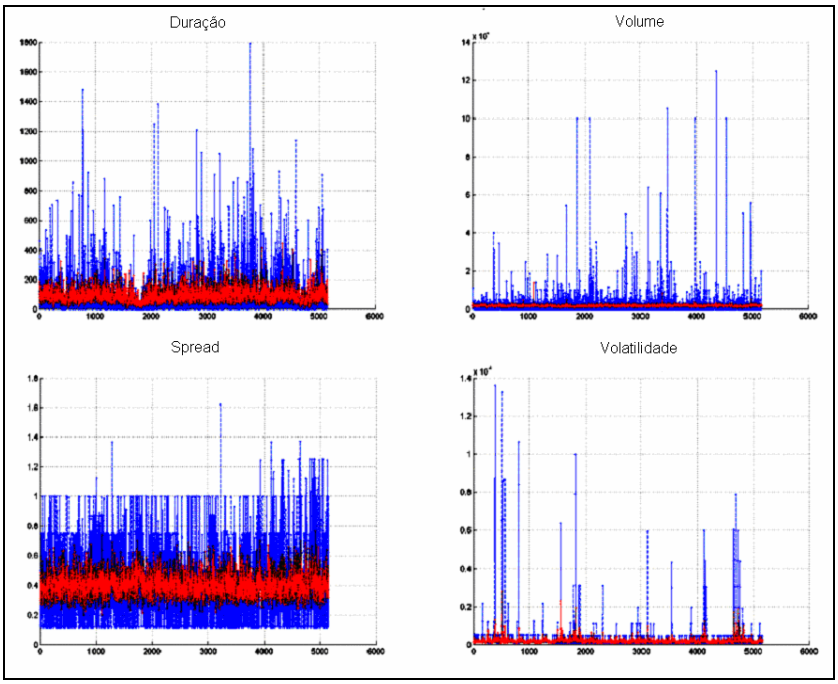

Figura 5.10: Previsto x realizado – variáveis financeiras de alta freqüência

- Comparação das estruturas propostas: a tabela 2 apresenta os resultados do teste de Ljung-Box. O teste toma por base os valores da função de autocorrelação. A hipótese nula estabelece que não existe dependência linear na série em estudo.

		Aceita H_0 (95%)	P-Valor	Ljung-Box	Valor Crítico
Observações	Duração	Rejeita	0,00%	442,25	25,00
	Volume	Rejeita	1,95%	28,34	25,00
	Spread	Rejeita	0,00%	774,45	25,00
	Volatilidade	Rejeita	0,00%	1486,72	25,00
Completo	Duração	Rejeita	0,00%	75,05	25,00
	Volume	Rejeita	1,30%	29,69	25,00
	Spread	Rejeita	0,00%	219,55	25,00
	Volatilidade	Rejeita	0,00%	206,89	25,00
"Livre de Variação"	Duração	Rejeita	0,08%	38,37	25,00
	Volume	Rejeita	1,17%	30,04	25,00
	Spread	Rejeita	0,00%	295,40	25,00
	Volatilidade	Rejeita	0,00%	118,14	25,00
Individual	Duração	Rejeita	0,00%	40,52	25,00
	Volume	Rejeita	15,88%	20,35	25,00
	Spread	Rejeita	0,00%	419,36	25,00
	Volatilidade	Rejeita	0,00%	620,99	25,00

Tabela 2: Resultados Ljung-Box – dependência linear

Com base nos resultados obtidos, pode-se observar que a formulação proposta reduz consideravelmente a dependência linear presente nas séries originais. Entretanto, conforme observado na figura 5.10, existe um excesso de dispersão nos resíduos relacionados a todas as variáveis em estudo. Isto se deve, provavelmente, às não-linearidades, conforme destacado por Engle e Russell (1998), Fernandes e Gramming (2001) e Zang, Russell e Tsay (2001).

A formulação restrita (*variation-free*) obteve melhor desempenho no que diz respeito ao ajuste dentro da amostra. De modo a testar a validade da adoção de restrições, foi utilizado o Teste da Razão de Verossimilhança. A hipótese nula estabelece a validade da formulação restrita quando comparada com a que

apresenta maior número de graus de liberdade. A tabela 3 apresenta os principais resultados.

	Aceita H_0 (95%)	P-Valor	Teste de Razão de Verossimilhança	Valor Crítico (95%)
Completo x "Livre de Variação"	Aceita	70,06%	27,11	36,42
Completo x Individual	Rejeita	0,00%	1123,87	72,15
"Livre de Variação" x Individual	Rejeita	0,00%	1096,76	43,77

Tabela 3: Resultados do Teste de Razão de Verossimilhança

Conforme proposto por Manganelli, a formulação “livre de variação” (*variation-free*) se mostra mais indicada do ponto de vista estatístico (p-valor = 70%). O teste de significância que considera a formulação individual rejeita fortemente a hipótese de que as variáveis apresentam dinâmicas independentes umas das outras.

Ainda no que diz respeito à adoção de restrições e tomando por base a formulação “livre de variação”, foi realizado um teste de razão de verossimilhança para ratificar a relação de causalidade imposta pela incorporação da matriz B_0 ao sistema. A tabela 4 apresenta os resultados.

Causalidade	Aceita H_0 (95%)	P-Valor	Teste de Razão de Verossimilhança	Valor Crítico (95%)
Duração → Volume	Rejeita	0,00%	606,29	3,84
Duração e Volume → Spread	Aceita	46,59%	1,53	5,99
Duração, Volume e Spread → Volatilidade	Rejeita	0,00%	371,72	7,81

Tabela 4: Teste de Razão de Verossimilhança (relação de causalidade)

Apesar dos resultados do teste apontarem para a não existência de qualquer tipo de interferência contemporânea da duração e do volume sobre a média condicional do *spread* de compra e venda, a imposição desta restrição compromete a boa aderência apresentada até então pelo modelo.