ANÁLISE BAYESIANA DE SENSIBILIDADE DO MODELO AR(1) PARA DADOS EM PAINEL: UMA APLICAÇÃO EM DADOS TEMPORAIS DE MICROARRAYS

Telma Suely da Silva MORAIS[1]
Fabyano Fonseca e SILVA[1]
Carlos Henrique Osório SILVA[1]
Sebastião MARTINS FILHO[1]
Moysés NASCIMENTO[1]
Thelma SÁFADI[2]

- **RESUMO:** Considerou-se uma análise Bayesiana do modelo auto-regressivo de primeira ordem, AR (1), para dados em painel, de forma a utilizar a função de verossimilhança exata, a análise de comparação de distribuições *a priori* (análise de sensitividade) e a obtenção de distribuições preditivas de dados futuros. A eficiência da metodologia proposta foi avaliada mediante um estudo de simulação, no qual a distribuição Beta re-escalada foi usada para representar 3 diferentes distribuições *a priori*: simétrica, assimétrica e uniforme. Realizou-se uma aplicação em dados reais de expressão gênica temporal de células HeLa gerados por microarray. Os resultados mostraram alta eficiência na previsão da expressão gênica para um instante futuro.

- KEYWORDS: Dados em painel; modelo auto-regressivo; inferência Bayesiana; *Microarray Time Series* (MTS).

1 Introdução

A área de Séries Temporais engloba uma gama enorme de modelos, dentre os quais um dos mais simples e utilizado é o modelo auto-regressivo de primeira ordem, AR (1). Ao generalizar este modelo para a análise simultânea de várias Séries Temporais tem-se o AR(1) para dados em painel, o qual considera informações de todas as séries para estimar parâmetros individuais de cada uma delas.

Em recentes estudos (SILVA *et al*. , 2008, SILVA *et al*., 2009) a metodologia Bayesiana foi utilizada com sucesso na análise de Séries Temporais para dados em painel, uma vez que as distribuições preditivas geraram valores acurados para observações em tempos futuros. Uma das razões deste sucesso foi a comparação de distribuições *a priori*,

[1] Universidade Federal de Viçosa - UFV, Centro de Ciências Exatas e Tecnológicas, Departamento de Estatística, Campus Universitário, CEP:36570-000, Viçosa, MG, Brasil. E-mail: *tel_morais@vicosa.ufv.br* / *fabyanofonseca@ufv.br* / *chos@dpi.ufv.br* / *martinsfilho@ufv.br* / *moysesnascim@gmail.com*

[2] Universidade Federal de Lavras - UFLA, Departamento de Ciências Exatas, Campus Universitário, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *safadi@ufla.br*

denominada de análise de sensitividade, a qual identificou as distribuições mais adequadas para representar os parâmetros de interesse.

Diante da relevância da análise de sensitividade, uma forma prática e eficiente de se trabalhar com diferentes distribuições *a priori* é adotar uma mesma distribuição com diferentes valores para os hiperparâmetros. Para tanto, deve-se utilizar distribuições maleáveis, capazes de assumir diferentes formas de acordo com as especificações de tais valores. Assim, a distribuição Beta apresenta-se como uma alternativa viável para esta finalidade, uma vez que sua f.d.p é relativamente simples e flexível (SILVA, 2000).

Pesquisas relacionadas com a aplicação de métodos Bayesianos em estudos de dados longitudinais (SAVIAN *et al*., 2009 e SILVA *et al*., 2009) utilizaram a técnica de simulação de dados para validar a teoria e os recursos computacionais adotados. Porém, faz-se necessário, principalmente no campo da Estatística Aplicada e Biometria, empregar os métodos desenvolvidos para solução de problemas relevantes e atuais. Dentre estes, destaca-se a análise de dados de Expressão Gênica temporal, denominados Microarray Time Series (MTS) (MUKHOPADHYAY e CHATTERJEE, 2007).

O objetivo deste trabalho foi propor uma análise Bayesiana do modelo AR(1) para dados em painel e comparar distribuições *a priori* caracterizadas por diferentes distribuições Beta (simétrica, assimétrica e uniforme). Objetivou-se ainda validar a eficiência da metodologia proposta via simulação de dados e aplicá-la em um conjunto de dados reais MTS.

2 Material e métodos

2.1 Modelo AR(1) para dados em painel

O modelo AR(1) para dados em painel é dado por (LIU e TIAO, 1980):

$$Z_{ij} = \phi_i Z_{i(j-1)} + e_{ij},\ i = 1,2,3,...,m; j = 1,2,3,...,n_i. \quad (1)$$

em que: Z_{ij} é o valor da observação da série i no tempo j, ϕ_i é o parâmetro de auto-regressão de cada série i e e_{ij} é o termo de erro aleatório, $e_{ij} \sim N(0,\sigma^2)$. Dessa forma, assume-se que $Z_{ij} \sim N(\phi_i Z_{i(j-1)}, \sigma^2)$. No que se segue, será considerado $n_1 = n_2 = ...= n_m =$ n, ou seja, o mesmo número de observações longitudinais para cada série i. A partir da distribuição de Z_{ij} é possível construir a função de verossimilhança, dada por $P(\mathbf{Z}_i \mid \phi_i, \sigma^2)$, a qual representa a distribuição conjunta dos dados de cada série. No presente trabalho, adotou-se a função de verossimilhança exata (2), uma vez que esta considera um termo referente à primeira observação de cada série, $P(\mathbf{Z}_{i1} \mid \phi_i, \sigma^2)$, e outro termo denominado de função de verossimilhança aproximada ou condicional, dada por $P(\mathbf{Z}_i^* \mid \phi_i, \sigma^2)$. Neste último, o índice j=1 não é considerado devido à defasagem observada no índice j do termo $Z_{i(j-1)}$. Assim, tem-se:

$$P(\mathbf{Z}_i \mid \phi_i, \sigma^2) = P(\mathbf{Z}_{i1} \mid \phi_i, \sigma^2) \times P(\mathbf{Z}_i^* \mid \phi_i, \sigma^2), \text{ sendo:} \quad (2)$$

$$P(\mathbf{Z}_{i1} \mid \phi_i, \sigma^2) \propto \left(\frac{1-\phi_i^2}{\sigma^2}\right)^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}\left(Z_{i1}^2\left(1-\phi_i^2\right)\right)\right\} \; e$$

$$P(\mathbf{Z}_i^* \mid \phi_i, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n-1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=2}^{n} Z_{ij}^2 - 2\phi_i \sum_{j=2}^{n} Z_{ij} Z_{i(j-1)} + \phi_i^2 \sum_{j=2}^{n} Z_{i(j-1)}^2\right\}.$$

Denominando:

$$h_i = \sum_{j=2}^{n-1} Z_{ij}^2, \; \hat{\phi}_i = \frac{\sum_{j=2}^{n} Z_{ij} Z_{i(j-1)}}{\sum_{j=2}^{n-1} Z_{ij}^2} \text{ e } S_i^2 = \sum_{j=1}^{n} Z_{ij}^2 - h_i \hat{\phi}_i^2, \text{ tem-se:}$$

$$P(\mathbf{Z}_i \mid \phi_i, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \times \left(1-\phi_i^2\right)^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}\left[S_i^2 + h_i\left(\hat{\phi}_i^2 - 2\phi_i \hat{\phi}_i + \phi_i^2\right)\right]\right\},$$

$$\text{ou } P(\mathbf{Z} \mid \boldsymbol{\phi}, \sigma^2) = \prod_{i=1}^{m} P(\mathbf{Z}_i \mid \phi_i, \sigma^2), \text{em que: } \mathbf{Z}=[\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_m], \; \mathbf{Z}_i=[\mathbf{Z}_{i1}, \mathbf{Z}_{i2}, ... \tag{3}$$

e $\boldsymbol{\phi} = [\phi_1, \phi_2, ..., \phi_m]$.

2.2 **Distribuições *a priori***

Optou-se por utilizar a distribuição Beta re-escalada no intervalo [-1,1] como distribuição *a priori* para ϕ_i, de forma a assumir diferentes valores para os seus parâmetros (α e λ). Foram utilizados valores que representassem as seguintes classes de distribuições: simétrica, assimétrica e uniforme. A expressão que representa a distribuição *a priori* para ϕ_i é a seguinte:

$$P(\phi_i|\alpha,\lambda) = \frac{1}{\beta(\alpha,\lambda)} \cdot \frac{1}{\left(1-(-1)\right)^{\alpha+\lambda-1}} \cdot \left(\phi_i - (-1)\right)^{\alpha-1} \cdot \left(1-\phi_i\right)^{\lambda-1}$$

.

Como a primeira parte desta expressão não contém o parâmetro de interesse ϕ_i, tem-se:

$$P(\phi_i|\alpha,\lambda) \propto \left(\phi_i + 1\right)^{\alpha-1} \left(1-\phi_i\right)^{\lambda-1} \tag{4}$$

Esta distribuição enfatiza um dos objetivos do presente trabalho, que consiste na utilização de uma única distribuição para representar várias outras com diferentes formas, de acordo com os valores assumidos para α e λ.

Para a variância residual a distribuição *a priori* adotada foi a Distribuição Gama-Inversa: $P(\sigma^2|c,d) \propto \left(\sigma^2\right)^{-(c+1)} \exp\left\{-\frac{d}{\sigma^2}\right\}$, em que c e d são os hiperparâmetros.

2.3 Distribuição conjunta *a posteriori* e distribuições condicionais completas *a posteriori*

De acordo com o teorema de Bayes, a distribuição conjunta *a posteriori* é:

$$P(\phi_i,\sigma^2|\mathbf{Z}_i,\alpha,\lambda,c,d) \propto P(\mathbf{Z}_i \mid \phi_i,\sigma^2)\times P(\phi_i \mid \alpha,\lambda)\times P(\sigma^2|c,d),$$

$$P(\phi_i,\sigma^2|\mathbf{Z}_i,\alpha,\lambda,c,d) \propto (\sigma^2)^{-\frac{n}{2}}(1-\phi_i^2)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[h_i(\phi_i^2-2\phi_i\hat{\phi}_i+\hat{\phi}_i^2)\right.\right. \quad (5)$$

$$\left.\left.(\phi_i+1)^{\alpha-1}(1-\phi_i)^{\lambda-1}\times(\sigma^2)^{-(c+1)} \exp\left\{\frac{d}{\sigma^2}\right\}\right.\right.$$

Para a implementação dos algoritmos MCMC, obteve-se a seguinte distribuição condicional completa *a posteriori* para ϕ_i:

$$P(\phi_i|\mathbf{Z}_i,\sigma^2) \propto (1-\phi_i)^{\lambda-\frac{1}{2}}(1+\phi_i)^{\alpha-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[h_i\left(\hat{\phi}_i^2-2\phi_i\hat{\phi}_i+\phi_i^2\right)\right]\right\}. \quad (6)$$

Nota-se que a expressão (6) não se caracteriza como uma distribuição de probabilidade conhecida, fato este que implica na utilização do algoritmo Metropolis-Hastings para gerar valores de ϕ_i que representam amostras de sua distribuição marginal. Ao se condicionar em relação a ϕ_i, como cada série i considerada tem a mesma variância residual σ^2, é necessário a utilização de um produtório para se condicionar em relação a todas as séries simultaneamente. Assim, tem-se:

$$P(\sigma^2|Z_i,\boldsymbol{\phi}) \propto \left(\sigma^2\right)^{\left(\frac{mn}{2}+cm+m-1+1\right)} \text{x } \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{m}\left(\frac{h_i(\phi_i-\hat{\phi}_i)^2+2d}{2}\right)\right\} \quad (7)$$

De acordo com o aspecto da expressão (7), é possível notar a f.d.p. de uma distribuição Gama-Inversa, cujos parâmetros são designados por:

$$c^* = \frac{mn}{2}+cm+m-1 \quad \text{e} \quad d^*=\sum_{i=1}^{m}\left(\frac{h_i(\phi_i-\hat{\phi})^2+2d}{2}\right).$$

Dessa forma, tem-se a seguinte distribuição condicional completa *a posteriori*: $\sigma^2 \mid Z_i,\phi_i \sim GI(c^*,d^*)$. Uma vez que se trata de uma distribuição conhecida, é possível utilizar o algoritmo Gibbs sampler para gerar amostras da distribuição marginal *a posteriori* de σ^2.

2.4 Distribuição preditiva

Sob o enfoque Bayesiano, uma observação futura é descrita por uma distribuição condicional aos dados passados, denominada distribuição preditiva. Esta distribuição é dada por:

$$P(Z_{i(n+1)} \mid Z_i) = \int_{\sigma_e^2}\int_{\phi_i} L(Z_{i|n+1}|\phi_i,\sigma^2,Z_i) \times P(\phi_i,\sigma^2|Z_i,\alpha,\lambda,c,d) d_{\phi_i} d_{\sigma^2},$$

em que:

$$L(\mathbf{Z}_{i|n+1}|\phi_i,\sigma^2,\mathbf{Z}_i) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[S_i^{*2} + h_i^*(\phi_i - \hat{\phi}_i^*)^2\right]\right\}, S_i^{*2} = Z_{i(n+1)}^2 - h_i^* \hat{\phi}_i^*, \quad h_i^* = Z_{in}^2$$

e $\hat{\phi}_i^* = Z_{i(n+1)}Z_{in} / h_i^*$. Assim: $S_i^{*2} = Z_{i(n+1)}^2 - \frac{Z_{in}^2 Z_{i(n+1)}^2 Z_{in}^2}{Z_{in}^4} = 0.$

$$\text{Portanto, } L(Z_{i(n+1)}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[Z_{in}^2\left(\phi_i - \frac{Z_{i(n+1)}Z_{in}}{Z_{in}^2}\right)^2\right]\right\}. \text{ O termo } P(\phi_i,\sigma^2|Z_i,\alpha,\lambda,c,d)$$

representa a distribuição conjunta *a posteriori* , expressão (5). Dessa forma, obtém-se:

$$P(\mathbf{Z}_{i(n+1)} \mid \mathbf{Z}_i) = \int_{\sigma_e^2}\int_{\phi_i} \exp\left\{-\frac{1}{2\sigma^2}\left[Z_{in}^2\left(\phi_i - \frac{Z_{i(n+1)}Z_{in}}{Z_{in}^2}\right)^2\right]\right\} \times (\sigma^2)^{-\frac{n}{2}}(1-\phi_i^2)^{\frac{1}{2}} \times$$

$$\exp\left\{-\frac{1}{2\sigma^2}\left[h_i(\phi_i^2 - 2\phi_i\hat{\phi}_i + \hat{\phi}_i^2)\right]\right\} \times (\phi_i+1)^{\alpha-1}(1-\phi_i)^{\lambda-1} \times (\sigma^2)^{-(c-1)} \exp\left\{\frac{d}{\sigma^2}\right\} d_{\phi_i} d_{\sigma_e^2}$$

A resolução da integral acima apresenta grande complexidade, fato este que justifica a utilização de métodos MCMC. Para tanto, utilizou-se:

$$P\left(Z_{i(n+1)} \mid Z_i, \sigma^2, \phi_i\right) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[Z_{in}^2\left(\phi_i - \frac{Z_{i(n+1)}Z_{in}}{Z_{in}^2}\right)^2\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[Z_{in}^2 \frac{\left(\phi_i Z_{in} - Z_{i(n+1)}\right)^2}{Z_{in}^2}\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[-(Z_{i(n+1)} - \phi_i Z_{in})\right]^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left[Z_{i(n+1)} - \phi_i Z_{in}\right]^2\right\}.$$

Pode-se notar que:

$$Z_{i(n+1)} \mid Z_i, \sigma^2, \phi_i \sim N\left(\phi_i Z_{in}, \sigma^2\right). \quad (8)$$

Se constatada a convergência dos algoritmos MCMC, pode-se assumir que o conjunto de valores gerados para esta distribuição Normal (8), provenientes de cada q iteração dos algoritmos MCMC, constitui a distribuição preditiva para um dado futuro. A estimativa da observação futura, dada pela média de todos os valores gerados pela distribuição Normal em questão (8), é dada por $\hat{Z}_{i(n+1)}$.

2.5 Algoritmos MCMC e comparação de modelos

Os algoritmos Gibbs Sampler e Metropolis-Hastings foram implementados matricialmente no software estatístico R (R DEVELOPMENT CORE TEAM, 2008). Considerou-se, tanto no estudo de simulação, como na aplicação aos dados reais, uma cadeia de 20.000 iterações, das quais a primeira metade foi eliminada ("*burn-in*"). A constatação final da convergência foi realizada por meio dos critérios de Geweke (1992) e de Raftery e Lewis (1992), ambos avaliados no pacote BOA ("*Bayesian Output Analysis*") do software R (SMITH, 2007).

No presente estudo foram considerados os modelos 1, 2 e 3, que correspondem, respectivamente, as distribuições *a priori* simétrica (Beta re-escalada com α=13,5 e λ=5), assimétrica (Beta re-escalada com α=4 e λ=2) e uniforme (Beta re-escalada com α=1 e λ=1). Tais distribuições assumidas para ϕ_i são apresnetadas na Figura 1.

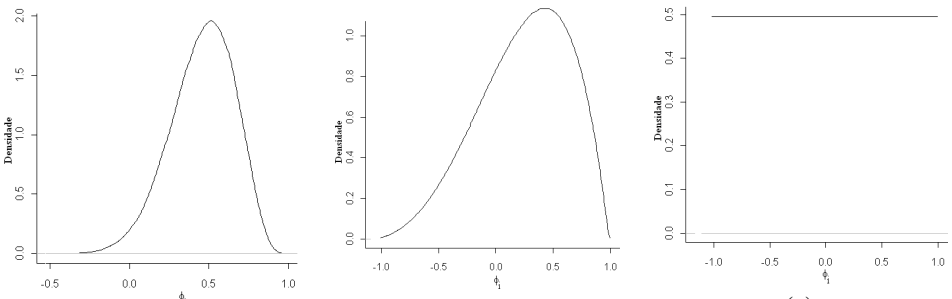

Figura 1 - Gráficos das f.d.p.´s para as distribuições *a priori* simétrica (a), assimétrica (b) e uniforme (c).

A comparação entre os modelos foi realizada via estimativa da probabilidade *a posteriori* do modelo (*posterior model probability*), a qual para um modelo ℓ em uma comparação envolvendo K modelos é dada por:

$$\hat{P}(M_\ell \mid \mathbf{Z}) = \frac{\hat{P}(\mathbf{Z} \mid M_\ell)P(M_\ell)}{\sum_{k=1}^{K} \hat{P}(\mathbf{Z} \mid M_k)P(M_k)}, \ k=1,2,...,\ell,...,K. \quad (9)$$

Na expressão (9), $\hat{P}(\mathbf{Z} \mid M_\ell)$ é a estimativa da Verossimilhança Marginal do modelo ℓ, e $P(M_\ell)$ corresponde ao peso *a priori* para este modelo, sabendo que $\sum_{k=1}^{K} P(M_k) = 1$. No presente estudo assumiu-se que $P(M_1) = ...P(M_2) = ... = P(M_\ell) = ... = P(M_K)$, e adotou-se:

$$\hat{P}(\mathbf{Z}|M_\ell) = \frac{1}{Q}\sum_{q=1}^{Q} P(\mathbf{Z}^{(q)}|\boldsymbol{\phi}^{(q)}, \sigma^{2(q)}) = \frac{1}{Q}\sum_{q=1}^{Q}\left(\prod_{i=1}^{m} P(Z_i^{(q)}|\phi_i^{(q)}, \sigma^{2(q)})\right),$$

$$\hat{P}(\mathbf{Z}|M_\ell) = \frac{1}{Q}\sum_{q=1}^{Q}\left[\prod_{i=1}^{m}\left(\frac{1}{\sigma^{2(q)}}\right)^{\frac{n}{2}} \times \left(1-\phi_i^{2(q)}\right)^{\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^{2(q)}}\left[S_i^2 + h_i\phi_i^2\left(\hat{\phi}_i^2 - 2\phi_i^{(q)}\hat{\phi}_i + \phi_i^{2(q)}\right)\right]\right\}\right],$$

o qual corresponde a média dos valores da função de verossimilhança (3) obtidos pela substituição dos valores atuais dos parâmetros gerados pelos algoritmos MCMC a cada q iteração (q = 1, 2,..., Q), sendo neste caso Q = 20.000.

Para avaliar a habilidade preditiva de cada modelo utilizou-se o recurso apresentado por Liu e Tiao (1980), o qual consiste na remoção da última observação de cada série. Assim, os parâmetros dos modelos são estimados sem a presença destas observações, as quais serão preditas pela metodologia adotada.

2.6 Dados simulados

A simulação de dados considerou a mesma configuração apresentada por Liu e Tiao (1980), e esta consta de 20 séries temporais (i=1, 2,...,20) cada uma com 8 observações longitudinais (j=1,2,...,8), mas como mencionado no item anterior, a última observação de cada série foi suprimida para avaliação da habilidade preditiva. A Figura 2 apresenta esquematização deste processo.

O procedimento apresentado na Figura 2 foi executado uma única vez para cada modelo, visto que a Inferência Bayesiana não necessita de repetições de experimentos para validar os resultados. Também vale ressaltar que na análise dos dados simulados, não se objetivou comparar os modelos 1, 2 e 3, e sim obter um resultado geral de validação de cada modelo separadamente. Para tanto, avaliou-se a qualidade de estimação dos parâmetros ϕ_i, σ^2 e Z_{i8} (estimativa da última observação, a qual foi excluída da análise). Para acessar a qualidade em questão, verificou-se se os valores paramétricos encontravam-se realmente contidos no intervalo de credibilidade de 95%. A comparação formal pela probabilidade *a posteriori* do modelo não foi considerada relevante no estudo de simulação, uma vez que ao simular valores de ϕ_i por uma distribuição simétrica, não foram encontradas razões para a qual uma distribuição assimétrica ou uniforme poderiam ser indicadas como a melhor.

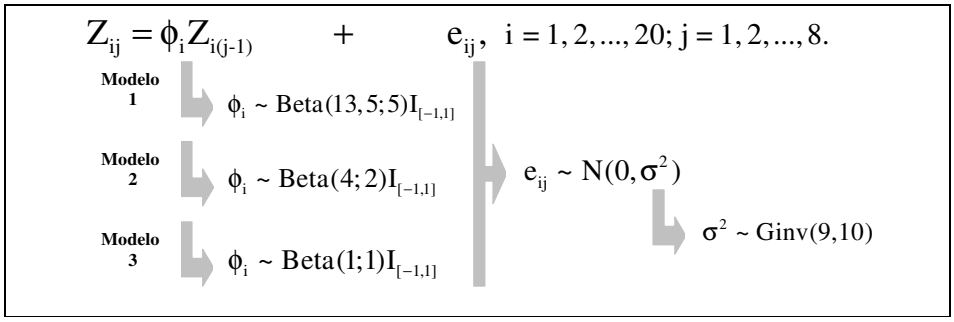

Figura 2 - Diagrama que descreve o processo de simulação dos parâmetros ao considerar cada um dos modelos considerados.

2.7 Dados reais

Sucintamente, a técnica de microarray consiste na deposição de seqüências de cDNA conhecidas em posições específicas de uma lâmina de vidro. Esta lâmina é dividida em vários quadrados, e cada um destes corresponde a um gene, isto é, em cada quadrado está fixada uma seqüência de DNA referente a um gene (FACELI *et al*., 2005; ESTEVES, 2007).

Suponha que se queira testar quais genes de um organismo atuam sobre a resistência às condições ambientais extremas. Deste modo, extrai-se o RNA total do organismo na presença (grupo tratado) e na ausência (grupo controle) das condições extremas. Após extrair o RNA, este é marcado por meio de técnicas de fluorescência. A próxima etapa é denominada de hibridização, e nesta as seqüências de RNA irão se juntar com as seqüências de DNA correspondentes (dispostos na lâmina). Posteriormente, essa lâmina é submetida a um processo de digitalização, que consiste na emissão de sinais luminosos que são captados pelo scanner, e quanto maior o sinal emitido pela fluorescência maior a expressão do gene. Como se sabem quais os genes correspondem a cada um dos quadrados na lâmina, é possível determinar quais genes estão sendo mais ou menos expressos em cada uma dos grupos (tratado e controle).

A variável resposta em um experimento de microarray é razão entre a intensidade de luz emitida pelos genes do grupo tratado e do grupo controle. Para facilidade de interpretação, toma-se o log2 desta razão, o qual é denominado *fold-change* (FC). Estes valores são calculados para cada gene, sendo que um valor de Fc positivo indica que o gene se expressa mais no grupo tratado (gene *up-regulated* para tratamento), um valor iguail a zero indica que o gene é igualmente expresso em ambos os grupos, e um valor negativo indica que o gene se expressa mais no grupo controle (gene *down-regulated* para tratamento). Os dados MTS (*microarray time series*) são caracterizados pela obtenção de valores FC ao longo do tempo, de forma que se torna possível avaliar o nível de expressão de cada gene em diferentes instantes.

Os dados utilizados no presente estudo são referentes à expressão de genes que atuam sobre células HeLa (células humanas epiteliais provenientes da fase final de crescimento) (WHITFIELD *et al*., 2002). Avaliou-se a expressão na presença (grupo tratado) e na ausência (grupo controle) de condições extremas de temperatura. Foram considerados no presente trabalho apenas oito genes (*reckp*, *cmycp*, *srcp*, *timp2p*, *ikkap*,

nfkbp, *nemop* e *nikp*) cuja relevância já foi relatada em outras pesquisas (FUJITA, 2007). Cada um destes genes apresentava 8 valores de FC, medidos em intervalos de uma em uma hora, sendo a última observação temporal de cada gene excluída para verificar a habilidade preditiva de cada modelo. Assim, o arquivo de dados resultante foi composto por oito séries, cada uma com sete observações. Todo o conjunto de dados utilizado está disponível no seguinte endereço eletrônico: http://genome-www.stanford.edu/Human-CellCycle/HeLa/.

3 Resultados e discussões

3.1 Dados simulados

As Tabelas 1, 2 e 3 apresentam as estimativas dos parâmetros do modelo AR(1) para dados em painel, respectivamente para cada distribuição *a priori* utilizada, ou seja, simétrica, assimétrica e uniforme.

Tabela 1 - Valores paramétricos (ϕ_i) simulados pelo modelo 1 (distribuição *a priori* simétrica), média da distribuição *a posteriori* ($\hat{\phi}_i$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Série	ϕ_i	$\hat{\phi}_i$	Li	Ls	P_G	FD_{RL}
1	0,561	0,569	0,560	0,576	0,571	0,993
2	0,477	0,480	0,475	0,483	0,717	0,985
3	0,191	0,192	0,189	0,193	0,797	1,006
4	0,399	0,393	0,390	0,406	0,569	0,985
5	0,696	0,688	0,690	0,698	0,746	1,001
6	0,256	0,260	0,250	0,265	0,566	0,993
7	0,256	0,258	0,243	0,261	0,034	0,961
8	0,630	0,635	0,623	0,640	0,839	1,026
9	0,890	0,897	0,887	0,905	0,603	1,018
10	0,488	0,486	0,480	0,489	0,610	1,010
11	0,195	0,197	0,191	0,199	0,806	0,977
12	0,718	0,725	0,708	0,732	0,406	1,001
13	0,211	0,204	0,200	0,257	0,889	1,001
14	0,619	0,610	0,603	0,639	0,974	0,977
15	0,762	0,759	0,734	0,798	0,405	1,026
16	0,650	0,653	0,640	0,656	0,760	1,026
17	0,845	0,848	0,840	0,851	0,661	1,052
18	0,368	0,373	0,361	0,378	0,861	1,001
19	0,524	0,532	0,521	0,539	0,293	1,052
20	0,483	0,482	0,477	0,498	0,604	1,060

Tabela 2 - Valores paramétricos (ϕ_i) simulados pelo modelo 2 (distribuição *a priori* assimétrica), média da distribuição *a posteriori* ($\hat{\phi}_i$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL}).

Série	ϕ_i	$\hat{\phi}_i$	Li	Ls	P_G	FD_{RL}
1	0,559	0,554	0,548	0,567	0,066	0,993
2	0,317	0,316	0,314	0,323	0,747	1,001
3	0,990	0,994	0,890	0,999	0,077	1,018
4	0,738	0,741	0,713	0,762	0,429	1,001
5	0,878	0,871	0,864	0,883	0,976	0,977
6	0,183	0,164	0,146	0,195	0,799	1,010
7	0,234	0,229	0,224	0,239	0,212	0,985
8	0,815	0,798	0,781	0,890	0,034	1,018
9	0,458	0,455	0,452	0,461	0,578	0,993
10	0,531	0,520	0,510	0,549	0,200	1,010
11	0,937	0,926	0,915	0,940	0,286	1,060
12	0,160	0,153	0,146	0,175	0,803	0,993
13	0,516	0,523	0,511	0,525	0,220	1,006
14	0,338	0,343	0,333	0,347	0,223	1,069
15	-0,010	-0,007	-0,019	-0,002	0,395	0,985
16	0,364	0,371	0,361	0,373	0,856	0,993
17	0,844	0,825	0,804	0,849	0,986	1,001
18	-0,115	-0,106	-0,151	-0,006	0,148	1,001
19	0,453	0,437	0,403	0,472	0,428	1,010
20	0,296	0,293	0,291	0,305	0,613	0,969

Tabela 3 - Valores paramétricos (ϕ_i) simulados pelo modelo 3 (distribuição *a priori* uniforme), média da distribuição *a posteriori* ($\hat{\phi}_i$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Série	ϕ_i	$\hat{\phi}_i$	Li	Ls	P_G	FD_{RL}
1	0,461	0,462	0,444	0,539	0,564	1,018
2	0,491	0,505	0,450	0,522	0,110	1,018
3	-0,106	-0,105	-0,134	-0,007	0,922	1,001
4	0,213	0,223	0,165	0,252	0,390	1,026
5	0,580	0,605	0,387	0,697	0,710	0,977
6	0,152	0,185	0,128	0,243	0,742	0,977
7	-0,100	-0,107	-0,154	-0,006	0,791	1,001
8	0,572	0,580	0,465	0,634	0,833	0,993
9	0,885	0,881	0,789	1,112	0,431	1,026
10	0,503	0,516	0,438	0,598	0,799	0,993
11	-0,203	-0,195	-0,237	-0,111	0,190	0,977
12	0,319	0,328	0,235	0,385	0,021	1,001
13	0,188	0,194	0,156	0,234	0,045	1,001
14	0,642	0,648	0,567	0,723	0,421	0,985
15	0,482	0,517	0,476	0,617	0,779	1,043
16	0,472	0,478	0,468	0,567	0,422	1,010
17	0,939	0,954	0,867	1,135	0,413	1,018
18	0,182	0,203	0,175	0,234	0,728	0,969
19	0,525	0,542	0,488	0,671	0,734	1,018
20	0,257	0,271	0,198	0,397	0,693	1,010

As Tabelas 4, 5 e 6 apresentam as estimativas para a última observação de cada série por meio da distribuição preditiva, respectivamente para cada distribuição utilizada como distribuição *a priori*, ou seja, simétrica, assimétrica e uniforme.

Tabela 4 - Última observação (Z_{i8}) simulada pelo modelo 1 (distribuição *a priori* simétrica), média da distribuição preditiva ($\hat{Z}_{i8}$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Série	Z_{i8}	$\hat{Z}_{i8}$	Li	Ls	P_G	FD_{RL}
1	-0,213	-1,148	-4,073	1,835	0,455	0,993
2	0,548	0,596	-2,385	3,574	0,207	1,001
3	-0,723	-0,097	-3,048	2,885	0,694	1,001
4	0,536	0,407	-2,532	3,389	0,703	1,035
5	-2,164	-1,705	-4,672	1,282	0,700	1,078
6	0,491	0,129	-2,833	3,037	0,139	1,010
7	-0,824	-0,293	-3,276	2,734	0,878	1,001
8	-0,460	-0,880	-3,934	2,152	0,837	0,969
9	-1,958	-1,485	-4,439	1,464	0,922	1,035
10	0,209	0,288	-2,660	3,341	0,586	1,010
11	-0,062	-0,042	-2,975	2,947	0,144	1,010
12	-2,122	-2,165	-5,109	0,757	0,082	1,001
13	-2,122	-0,227	-3,177	2,715	0,229	0,977
14	0,353	0,339	-2,643	3,263	0,693	1,001
15	0,357	0,345	-2,562	3,265	0,658	1,001
16	2,239	1,404	-1,507	4,308	0,554	0,993
17	-0,128	0,277	-2,658	3,213	0,553	1,001
18	-0,453	0,122	-2,812	3,062	0,341	0,993
19	1,388	0,741	-2,203	3,641	0,737	1,001
20	1,488	1,139	-1,778	4,121	0,822	0,993

Tabela 5 - Última observação (Z_{i8}) simulada pelo modelo 2 (distribuição *a priori* assimétrica), média da distribuição preditiva ($\hat{Z}_{i8}$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Série	Z_{i8}	$\hat{Z}_{i8}$	Li	Ls	P_G	FD_{RL}
1	0.332	0.744	-2.710	4.260	0.589	1.001
2	-0.097	0.335	-3.156	3.769	0.094	1.026
3	-1.930	-0.663	-4.094	2.732	0.245	1.035
4	2.199	0.203	-3.256	3.680	0.217	0.977
5	1.440	0.721	-2.673	4.147	0.072	0.993
6	1.682	0.375	-2.994	3.812	0.310	1.001
7	1.902	0.454	-2.942	3.851	0.867	0.993
8	-1.020	-2.942	-6.365	0.547	0.082	0.993
9	-1.501	-0.129	-3.531	3.326	0.210	1.060
10	-2.211	-0.129	-3.583	3.483	0.336	1.035
11	2.340	3.056	-0.429	6.488	0.203	0.993
12	-0.411	-0.081	-3.628	3.368	0.937	1.010
13	0.936	0.290	-3.174	3.777	0.937	1.014
14	-0.863	0.100	-3.317	3.612	0.386	0.993
15	0.441	-0.013	-3.435	3.465	0.475	1.018
16	-0.107	0.316	-3.089	3.813	0.450	0.993
17	3.271	1.205	-2.274	4.690	0.341	0.977
18	-0.599	-0.056	-2.274	3.438	0.641	0.985
19	-0.032	-0.350	-3.745	3.162	0.356	0.977
20	-1.578	0.158	-3.382	3.590	0.562	1.001

Tabela 6 - Última observação (Z_{i8}) simulada pelo modelo 3 (distribuição *a priori* uniforme), média da distribuição preditiva ($\hat{Z}_{i8}$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Série	Z_{i8}	$\hat{Z}_{i8}$	Li	Ls	P_G	FD_{RL}
1	1,465	0,074	-2,638	2,831	0,564	1,018
2	0,595	-0,019	-2,840	2,752	0,110	1,018
3	0,489	-0,241	-3,006	2,512	0,922	1,001
4	-1,054	0,104	-2,625	2,869	0,390	1,026
5	0,340	-0,268	-3,020	2,423	0,710	0,977
6	0,293	-1,228	-3,916	1,479	0,742	0,977
7	-0,772	0,002	-2,769	2,821	0,791	1,001
8	-1,326	0,949	-1,008	3,705	0,833	0,993
9	0,800	0,132	-2,600	2,797	0,431	1,026
10	0,347	-0,316	-3,068	2,389	0,799	0,993
11	0,046	0,149	-2,660	2,927	0,190	0,977
12	-0,417	-0,766	-3,601	1,998	0,021	1,001
13	0,142	-0,188	-2,969	2,617	0,045	1,001
14	-0,635	0,394	-2,337	3,162	0,421	0,985
15	-2,784	3,162	-3,674	1,859	0,779	1,043
16	1,827	1,744	-0,967	4,449	0,422	1,010
17	-0,416	-0,503	-3,254	2,220	0,413	1,018
18	1,924	0,485	2,287	3,196	0,728	0,969
19	-0,176	0,245	-2,532	2,963	0,734	1,018
20	-1,930	0,190	-2,573	2,909	0,693	1,010

De forma geral, todas as distribuições *a priori* utilizadas forneceram bons resultados em relação à predição de um único valor futuro, principalmente a distribuição *a priori* simétrica (Tabela 4), a qual apresentou uma habilidade preditiva de 100%, ou seja, para todas as 20 séries os intervalos de credibilidade continham o verdadeiro valor da última observação. É bom recordar que o mesmo é conhecido, pois foi simulado juntamente com os outros valores, mas não foi considerado na análise. A distribuição *a priori* assimétrica (Tabela 5) e a distribuição *a priori* uniforme (Tabela 6) apresentaram, respectivamente, capacidades preditivas de 80 e 90%, pois na Tabela 5 para quatro séries (4, 5,6 e 7) e na Tabela 6 para duas séries (8 e 18) os valores paramétricos não estavam inclusos no intervalo.

Estes resultados não nos permitem admitir que a distribuição *a priori* simétrica é o melhor modelo, pois não é apresentado um método formal de comparação de modelos, mesmo porque o conjunto de dados é diferente, ou seja, a distribuição *a priori* simétrica foi usada em uma situação em que os valores dos parâmetros foram gerados por uma distribuição simétrica, no caso Beta (13,5;5). Resumidamente, os resultados das Tabelas 4, 5 e 6 indicam que a distribuição preditiva apresenta-se como uma ferramenta eficiente para a predição de valores futuros, visto que ao se considerar os resultados de todas estas tabelas, tem-se uma habilidade preditiva de 90%. Silva *et al*. (2008) considerando as distribuições *a priori* t-Student multivariada, Normal-Multivariada e *priori* de Jeffreys obtiveram resultados semelhantes, com eficiência de 80%.

A Tabela 7 apresenta as estimativas da variância residual para cada distribuição *a priori* considerada.

Tabela 7 - Variância residual paramétrica (σ^2), média da distribuição *a posteriori* ($\hat{\sigma}^2$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Modelo (dist. *a priori*)	σ^2	$\hat{\sigma}^2$	Li	Ls	P_G	FD_{RL}
1 (dist. Beta simétrica)	1,8207	2,2567	1,2742	3,9379	0,8478	0,9938
2 (dist. Beta assimétrica)	2,3978	3,1094	1,7475	5,4065	0,3775	1,0018
3 (dist. Uniforme)	1,1494	1,9220	1,0886	3,3370	0,9674	0,9698

Quanto às estimativas apresentadas na Tabela 7, observa-se que para todas as distribuições *a priori* utilizadas, a variância estimada foi superior ao valor paramétrico, porém bem próximas deste valor. Observa-se ainda que analogamente aos resultados das Tabelas 1 a 3, todos os intervalos de credibilidade contiveram os valores paramétricos, indicando que a metodologia foi eficiente para estimar a variância residual. As estatísticas P_G e FD_{RL} informam que o número de iterações utilizadas foi suficiente para garantir a convergência das cadeias geradas pelo algoritmo Gibbs Sampler.

3.2 Dados reais

A Tabela 8 apresenta as estimativas dos parâmetros do modelo AR (1) para dados em painel, respectivamente para cada distribuição utilizada como *priori* (simétrica, assimétrica e uniforme) ajustada ao mesmo conjunto de dados de expressão gênica avaliada ao longo do tempo.

De acordo com os resultados apresentados na Tabela 8 nota-se que a distribuição *a priori* simétrica e a distribuição *a priori* uniforme forneceram uma porcentagem de significância um pouco menor que aquela obtida para a distribuição *a priori* assimétrica, uma vez que para esta última o intervalo de credibilidade não conteve zero para nenhuma das séries. Isto indica que ao se utilizar estas duas distribuições *a priori*, o modelo AR (1) não é eficiente para descrever a série 7 (gene nemop), podendo esta ser considerada um processo aleatório, que indica a ausência de autocorrelação, ou ser descrita por uma ordem superior, por exemplo, AR (2).

Outro ponto relevante quanto às estimativas mostradas na Tabelas 8 diz respeito ao sinal do parâmetro de auto-regressão, o qual foi negativo para todas as séries. Este fato significa que as observações da série de expressão de um gene em um dado instante t tendem a ser negativamente correlacionadas com as observações em um instante t-1, ou seja, se um gene apresenta grande expressão em um instante, no instante anterior ele apresentou uma expressão menor.

Tabela 8 - Média *a posteriori* ($\hat{\phi}_i$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Modelo 1 (distribuição *a priori* simétrica)					
Gene	$\hat{\phi}_i$	Li	Ls	P_G	FD_{RL}
1	-0,3310	-0,3929	-0,3129	0,7686	2,1964
2	-0,3984	-0,4203	-0,3803	0,6928	2,1190
3	-0,4171	-0,4791	-0,4391	0,5754	2,1601
4	-0,7280	-0,7699	-0,7599	0,5955	2,1393
5	-0,8595	-0,8915	-0,8515	0,2433	1,9685
6	-0,4543	-0,4963	-0,4963	0,0428	2,2557
7	-0,2808	-0,5528	0,03128	0,1311	2,0298
8	-0,3844	-0,4264	-0,3564	0,3318	2,1601
Modelo 2 (distribuição *a priori* assimétrica)					
Gene	$\hat{\phi}_i$	Li	Ls	P_G	FD_{RL}
1	-0,3244	-0,3814	-0,3129	0,2129	1,8376
2	-0,3814	-0,4645	-0,3803	0,0402	2,1740
3	-0,4123	-0,4451	-0,4391	0,1805	1,9685
4	-0,6976	-0,7866	-0,7599	0,2396	2,1131
5	-0,8342	-0,9234	-0,8515	0,2047	1,8280
6	-0,4443	-0,5005	-0,4963	0,3816	2,2584
7	-0,3228	-0,5228	-0,0128	0,9341	2,0080
8	-0,3356	-0,4689	-0,2674	0,0654	2,1665
Modelo 3 (distribuição *a priori* uniforme)					
Gene	$\hat{\phi}_i$	Li	Ls	P_G	FD_{RL}
1	-0,3219	-0,3804	-0,2804	0,0926	1,0184
2	-0,3911	-0,4696	-0,3496	0,2563	1,0101
3	-0,4146	-0,4732	-0,3532	0,0365	1,0523
4	-0,7237	-0,7622	-0,6922	0,4532	1,0018
5	-0,8477	-0,9263	-0,6863	0,4124	0,9855
6	-0,4511	-0,4897	-0,4897	0,4848	1,0437
7	-0,2704	-0,5390	0,0451	0,2107	0,9775
8	-0,3743	-0,4429	-0,3129	0,3525	1,0184

A descrição deste sistema envolve mecanismos de retroalimentação negativa, ou seja, os genes com grande expressão produzem grande quantidade de proteínas, e quando

este excesso está na presença de determinados receptores, é possível a formação de complexos (oligômeros) que são transportados do citoplasma para o núcleo, e esta presença no núcleo bloqueia sua própria transcrição ao inibir a ação de fatores de transcrição. De qualquer forma, esta inibição ainda caracteriza-se como um processo pouco esclarecido (MARTINS, 2007)

A Tabela 9 apresenta as estimativas para a última observação de cada série por meio da distribuição preditiva, respectivamente para cada distribuição *a priori* utilizada.

Tabela 9 - Valor verdadeiro observado (Z_{i8}), média da distribuição preditiva ($\hat{Z}_{i8}$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Modelo 1 (distribuição *a priori* simétrica)						
Gene	Z_{i8}	$\hat{Z}_{i8}$	Li	Ls	P_G	FD_{RL}
1	0,0510	0,0257	-2,9745	2,9491	0,9601	0,9938
2	0,6510	0,2916	-2,2867	2,7475	0,9749	0,9698
3	-0,0970	-0,089	-3,0150	2,9441	0,6509	1,0437
4	0,1830	0,2926	-3,3139	2,6872	0,5030	1,0018
5	0,0890	-0,0290	-2,1097	2,8597	0,6990	1,0104
6	0,1660	0,0968	-2,1971	2,9057	0,0372	1,0352
7	0,2030	0,1549	-3,1377	2,9000	0,5994	1,0018
8	0,0810	-0,01110	-2,1432	2,9056	0,0350	1,0101
Modelo 2 (distribuição *a priori* assimétrica)						
Gene	Z_{i8}	$\hat{Z}_{i8}$	Li	Ls	P_G	FD_{RL}
1	0,0510	0,0555	-2,0429	2,8882	0,3569	1,0184
2	0,6510	0,3194	-2,3283	2,5866	0,6171	0,9855
3	-0,0970	0,0114	-2,9896	2,9994	0,6977	1,0101
4	0,1830	0,3416	-2,2667	2,6288	0,5111	0,9855
5	0,0890	-0,0176	-2,0862	2,7890	0,7208	1,0101
6	0,1660	0,0682	-2,9435	2,8092	0,5436	0,9983
7	0,2030	0,1389	-3,0008	2,9978	0,7621	1,0018
8	0,0810	-0,0090	-2,0432	2,9422	0,1065	1,0266
Modelo 3 (distribuição *a priori* uniforme)						
Gene	Z_{i8}	$\hat{Z}_{i8}$	Li	Ls	P_G	FD_{RL}
1	0,0510	0,0427	-2,9488	2,8947	0,7501	0,9855
2	0,6510	0,3169	-3,2942	2,5888	0,4237	0,9855
3	-0,0970	0,0126	-2,9005	2,9789	0,5241	1,0440
4	0,1830	0,2768	-3,2319	2,7102	0,9934	1,0184
5	0,0890	-0,0183	-3,1620	2,7863	0,4525	0,9855
6	0,1660	0,0725	-3,1064	2,9557	0,3598	0,9855
7	0,2030	0,3694	-3,0466	2,9392	0,1128	1,0352
8	0,0810	-0,0106	-3,0245	2,7829	0,5105	1,0018

Os resultados apresentados na Tabela 9 permitem afirmar que a metodologia utilizada para realizar previsões de dados futuros de séries individuais com base na obtenção das distribuições preditivas foi eficiente, uma vez que a porcentagem de intervalos de credibilidade que continham os verdadeiros valores das expressões gênicas no último instante de tempo foi de 100% para todas as distribuições *a priori* utilizadas.

Na prática, a aplicação desta metodologia de previsão em dados de microarray avaliados ao longo do tempo apresenta-se como uma inovação tecnológica que permite predizer o valor da expressão gênica em tempos não estudados, reduzindo assim os custos relacionados com os procedimentos laboratoriais, os quais segundo Faceli (2005) são bastante significativos e até limitantes quanto à implantação de projetos na área de microarray. Neste caso, a redução dos custos seria caracterizada pela utilização do valor predito da expressão gênica em um dado tempo futuro não estudado, em vez da utilização do valor obtido de amostras avaliadas laboratorialmente neste mesmo tempo.

De forma geral os valores estimados apresentaram-se bem próximos dos verdadeiros valores para a última observação, porém esta análise pontual não apresenta grande relevância, uma vez que toda discussão foi efetuada de acordo com a estimação intervalar. Assim, com o intuito de explorar melhor estes resultados, confeccionou-se um gráfico (Figura 3) contendo as amplitudes dos intervalos de credibilidade de 95% para cada distribuição *a priori* adotada.

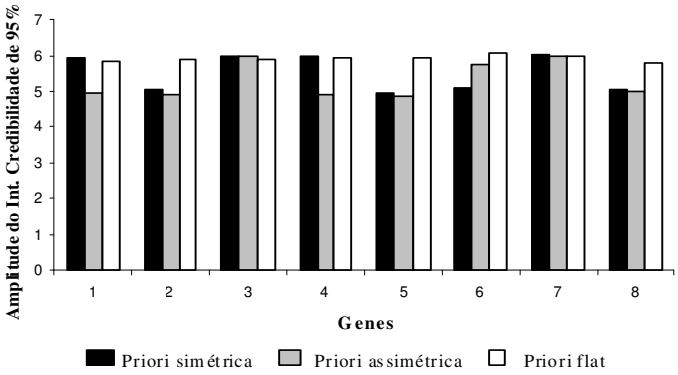

Figura 3 - Amplitude do intervalo de credibilidade de 95% para a última observação de cada série correspondente a expressão gênica de células HeLa.

A Tabela 10 apresenta as estimativas da variância residual para cada distribuição *a priori* considerada.

De acordo com a Figura 3 é possível visualizar que a utilização da distribuição *a priori* assimétrica proporcionou intervalos de credibilidade mais estreitos para cinco (genes 1, 2, 4, 5 e 8) das oito séries estudadas. Alheio a isto, nota-se na Tabela 10 que a estimativa da variância residual para esta mesma distribuição *a priori* foi inferior que as das demais, embora esta diferença seja de pequena magnitude. Estes resultados mostram indícios de uma maior qualidade da distribuição *a priori* assimétrica, porém torna-se relevante ressaltar que, embora estes apontem para esta distribuição como sendo a melhor,

segundo Silva *et al.* (2008), a avaliação da qualidade dos modelos, ou seja, das diferentes distribuições *a priori*, deve ser discutida mediante a utilização de um critério específico de comparação, como por exemplo, a probabilidade *a posteriori* do modelo (Tabela 11).

Tabela 10 - Estimativa dada pela média da distribuição *a posteriori* ($\hat{\sigma}^2$), limite inferior (Li) e superior (Ls) do intervalo de credibilidade de 95%, valor de probabilidade de Geweke (P_G) e Fator de dependência de Raftery e Lewis (FD_{RL})

Modelo (dist. *a priori*)	$\hat{\sigma}^2$	Li	Ls	P_G	FD_{RL}
1 (dist. Beta simétrica)	2,5911	0,9303	5,3317	0,1339	1,0101
2 (dist. Beta assimétrica)	2,1582	0,9219	5,3534	0,0741	1,0266
3 (dist. Uniforme)	2,6467	0,9445	5,3087	0,4439	1,0184

Tabela 11 - Comparação das distribuições *a priori* por meio do Fator de Bayes considerando os dados reais

Modelo (distribuições *a priori*)	Probabilidade *a posteriori* do modelo
1. (dist. Beta simétrica)	$\hat{P}(M_1 \mid \mathbf{Z}) = \frac{\hat{P}(\mathbf{Z} \mid M_1)P(M_1)}{\sum_{k=1}^{3} \hat{P}(\mathbf{Z} \mid M_k)P(M_k)} = 0,3277$
2. (dist. Beta assimétrica)	$\hat{P}(M_2 \mid \mathbf{Z}) = \frac{\hat{P}(\mathbf{Z} \mid M_2)P(M_2)}{\sum_{k=1}^{3} \hat{P}(\mathbf{Z} \mid M_k)P(M_k)} = 0,3377$
3. (dist. Uniforme)	$\hat{P}(M_3 \mid \mathbf{Z}) = \frac{\hat{P}(\mathbf{Z} \mid M_3)P(M_3)}{\sum_{k=1}^{3} \hat{P}(\mathbf{Z} \mid M_k)P(M_k)} = 0,3347$

Os resultados contidos na Tabela 11 refletem a indiferença das distribuições *a priori* utilizadas, possibilitando afirmar que todas as distribuições de probabilidade utilizadas apresentaram a mesma qualidade.

Os gráficos contendo os valores observados e estimados para cada série de expressão gênica, ao considerar a distribuição *a priori* assimétrica (maior valor de probabilidade *a posteriori* segundo Tabela 11), são apresentados na Figura 4. Nota-se nestes gráficos que, de forma geral, o modelo utilizado descreveu corretamente o comportamento da expressão de cada gene ao longo do tempo, uma vez que as tendências de crescimento e decrescimento foram, na maioria das vezes, respeitadas.

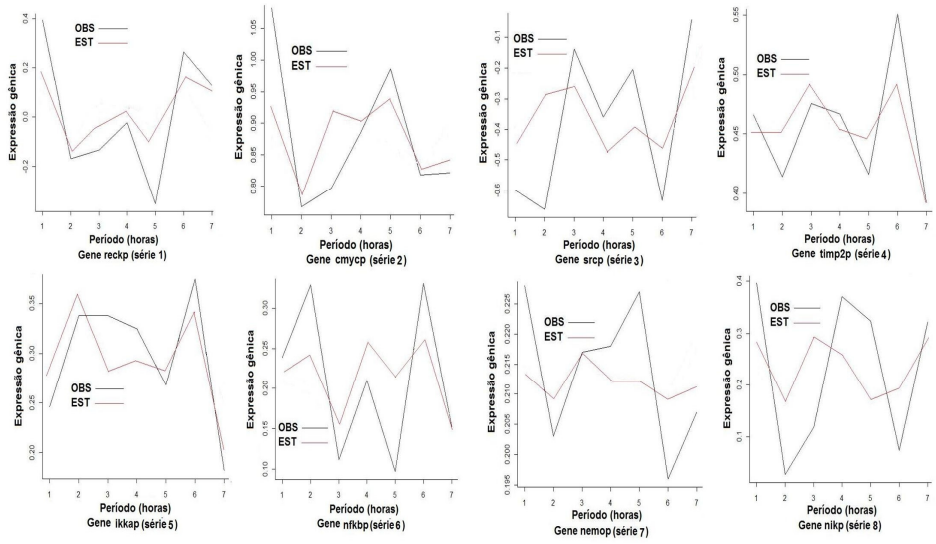

Figura 4 - Séries temporais de expressão gênica observadas (Obs) e estimadas (Est) pelo modelo auto-regressivo de primeira ordem, AR (1), para dados em painel considerando a distribuição *a priori* Beta assimétrica.

Conclusões

1) Os resultados do estudo de simulação indicaram que a metodologia proposta foi eficiente, pois na grande maioria das séries simuladas os verdadeiros valores dos parâmetros (ϕ_1,σ^2 e Z_{i8}) encontravam-se dentro do intervalo de credibilidade de 95%, indicando assim alta eficiência no processo de estimação.

2) A aplicação da metodologia proposta aos dados reais de microarray de células HeLa avaliados ao longo do tempo possibilitou informar o comportamento da expressão por meio das estimativas do parâmetro ϕ_1 para cada gene estudado, além disto a comparação de modelos indicou numericamente uma superioridade da distribuição *a priori* assimétrica, porém em termos práticos, devido à pequena diferença entre estes valores, pode-se afirmar que as três distribuições *a priori* utilizadas apresentaram a mesma qualidade.

JANGARELLI, M.; EUCLYDES, R. F.; CECON, P. R. Bayesian approach of AR(1) panel data model: Application in microarray time series data. *Rev. Bras. Biom.*, São Paulo, v.28, n.4, p.171-192, 2010.

▪ *ABSTRACT: We considered a Bayesian analysis of first order autoregressive, AR(1), panel data model, using exact likelihood function, comparative analysis of prior distributions (sensitivity*

analysis) and predictive distributions of future observations. The methodology efficiency was evaluated by a simulation study using three prior, which were related to different Generalized Beta distributions: symmetric, asymmetric and uniform prior. We applied the proposed methodology to microarray time series real data of HeLa cells. The forecast of gene expression in one future time showed high efficiency.

- *KEYWORDS: Panel data; autoregressive model; Bayesian inference; microarray time series.*

Referências

BARRETO, G.; ANDRADE, M. G. Robust bayesian approach for AR(p) models applied to streamflow forecasting. *Journal Applied Statistical Science*, New York, v.12, n.3, p.269-292, 2004.

ESTEVES, G. H. *Metodos estatısticos para a análise de dados de cDNA microarray em um ambiente computacional integrado*. 2008. 174 f. Tese (Doutorado em Bioinformática) – Universidade de São Paulo, São Paulo, 2008.

FACELI, K.; CARVALHO, A. C. P. L. F.; SOUTO, M.C.P. *Análise de dados de expressão gênica*, São Carlos, SP: ICMC, 2005. (Relatório técnico 250)

FUJITA, A. ; SATO, J. R.; FERREIRA, C. E.; SOGAYAR, M. C. GEDI: a user-friendly toolbox for analysis of large-scale gene expression data. *BMC Bioinformatics*, v.8, p.457, 2007.

GEWEKE, J. Evaluating the accurary of sampling-based approaches to the calculation of posterior moments. In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A. P.; SMITH, A. F. M. (Eds.). *Bayesian statistics*. New York, USA: Oxford University Press, 1992. 625-631p.

LIU, L. M.; TIAO, G. C. Randon coefficient first-order autoregressive model. *Journal of Econometrics***,** New York, v 13, n.3, p.305-325, 1980.

MARTINS, C.B.L.B. *Expressão gênica de melanopsina, clock.per e cry e sua modulação por melatonina em células de Danio Rerio***.** 2007. 60f. Dissertação (Mestrado) – Universidade de São Paulo, 2007.

MUKHOPADHYAY, M.; CHATTERJEE, S. Causality and pathway search in microarray time series experiment. *Bioinformatics*, v.23, p.442-449, 2007.

R DEVELOPMENT CORE TEAM. 2008. *R: a language and environment for statistical computing***.** Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: 2008.

RAFTERY, A.E.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J.M. et al. (Eds.). *Bayesian statistics*. Oxford, USA: University Press, 1992. p.763-773.

SAVIAN, T. V.; MUNIZ, J. A. ; SÁFADI, T. ; SILVA, F. F. . Análise bayesiana para modelos de degradabilidade ruminal. *Ciência Rural*, v.39, p.1752, 2009.

SILVA, F. F.; SÁFADI, T.; MUNIZ, J. A. ; AQUINO, L. H. ; MOURÃO, G. B. Comparação bayesiana de modelos de previsão de diferenças esperadas nas progênies no melhoramento genético de gado Nelore. *Pesquisa Agropecuária Brasileira*, v.43, p.37, 2008.

SILVA, F. F. ; SÁFADI, T. ; MUNIZ, J. A. ; AQUINO, L. H. ; MOURÃO, G. B. Previsão Bayesiana de valores genéticos de touros por meio do modelo auto-regressivo para dados em painel. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*, v.60, p.1166, 2009.

SILVA, C. H. O. *A proposed framework for establishing optimal genetic designs for estimating narrow-sense heritability*. 2000. 98f. Tese (Doutorado em Estatística) – North Carolina State University, NCSU, USA, 2000.

SMITH, B. J. Boa: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, v.21, n.11, p.1-37, 2007.

WHITFIELD, M. L.; SHERLOCK, G.; SALDANHA, A. J.; MURRAY, J. I.; BALL, C. A.; ALEXANDER, K. E.; MATESE, J. C.; PEROU, C. M.; HURT, M. M.; BROWN, P. O.; BOTSTEIN, D. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, v. 13, p.1977-2000, 2002.

Recebido em 21.06.2010.

Aprovado após revisão 21.09.2010.