TELMA SUELY DA SILVA MORAIS

ABORDAGEM BAYESIANA DO MODELO AR(1) PARA DADOS EM PAINEL: UMA APLICAÇÃO EM DADOS TEMPORAIS DE MICROARRAY

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de Magister Scientiae.

VIÇOSA
MINAS GERAIS – BRASIL
2008

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
M827a
2008

Morais, Telma Suely da Silva, 1971-
Abordagem Bayesiana do modelo AR(1) para dados em painel : uma aplicação em dados temporais de microarray / Telma Suely da Silva Morais. – Viçosa, MG, 2008.
ix, 46f. : il. (algumas col.) ; 29cm.

Inclui apêndice.
Orientador: Fabyano Fonseca e Silva.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 50-54.

1. Teoria bayesiana de decisão estatística. 2. Análise de séries temporais. 3. Inferência (Lógica). 4. Estatística matemática. 5. Probabilidades. I. Universidade Federal de Viçosa. II.Título.

CDD 22.ed. 519.342

TELMA SUELY DA SILVA MORAIS

ABORDAGEM BAYESIANA DO MODELO AR(1) PARA DADOS EM PAINEL: UMA APLICAÇÃO EM DADOS TEMPORAIS DE MICROARRAY

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de Magister Scientiae.

APROVADA: 05 de dezembro de 2008.

Carlos Henrique Osório Silva
(Co-Orientador)

Paulo Roberto Cecon
(Co-Orientador)

Sebastião Martins Filho

Sidney Martins Caetano

Fabyano Fonseca e Silva
(Orientador)

Aos meus pais Antônio e Solange (in memoriam), fonte de dignidade e amor incondicional.

Aos meus irmãos Cris, Cinha, Ninha e Binho, fontes de carinho, união e cumplicidade.

Aos meus cunhados Otávio e Karol, símbolos de amizade, carinho e incentivo.

Às minhas sobrinhas Yasmin, Beatriz e Maria Clara, luzes da minha vida.

Aos os meus familiares, exemplos de apoio constante.

Dedico!

"A mente que se abre a uma nova idéia jamais voltará ao seu tamanho original."

(Einstein)

AGRADECIMENTOS

Aos meus pais Antônio e Solange (in memoriam), por todos os valores transmitidos a mim e pelo apoio incondicional em toda minha caminhada.

Aos meus irmãos, Cris, Cinha, Ninha e Binho, pelo incentivo, apesar desse sonho ter nos roubado o convívio.

À Universidade Federal de Viçosa, particularmente ao Departamento de Informática, Setor de Estatística, pela oportunidade de realização do curso.

Ao professor Fabyano, pela orientação e grandeza da sua generosidade e amizade e pela confiança depositada em mim.

Aos professores Sebastião, Carlos Henrique, Peternelli, Policarpo e José Ivo, pelos ensinamentos repassados.

Ao professor Cecon, pela amizade e agradável convívio.

Às professoras Marinês (DMA), Margareth (DMA) e Tânia (EDU) por terem acreditado em mim.

Ao secretário da Pós-Graduação Altino, pelo empenho em sempre ajudar e pelas mensagens de amizade e incentivo.

Às funcionárias Eliana e Marisa, pela colaboração.

Ao funcionário Paulinho, pelos cafezinhos e pelo sorriso no rosto.

À minha colega de curso e amiga Flávia, por ter aturado as minhas lamentações e por ter dividido comigo seus conhecimentos.

À Fê, pela colaboração com seus conhecimentos para a realização desse trabalho (Estatística Genética é linda!).

Aos meus queridos colegas Alex, Carol, Moisés, Andréia, Thiago, Braúlia, Danilo, Willerson, Nilsa e Toninho (Enciclopédia), pelos momentos de descontração, companheirismo e amizade.

Às minhas amigas Renata, Neti, Déia, Kátia, Carol, Evânia,Rogéria, Mari, Lé, Ika, Jô, Rosi e Claudinha, por estarem sempre presentes na minha vida.

A Josy e Raquel, pela amizade e por terem-me acolhido quando eu precisava.

Às minhas companheiras de república Carla, Bruna, Chris e Fê, pelo incentivo e amizade.

Ao meu amigo Oscar, por me fazer crer que algumas coisas são mesmo para sempre.

À minha "irmã" Laura (in memoriam), por ter sido tão especial e ter-me feito crer que todos temos mesmo uma missão a cumprir.

A cidade e população de Viçosa, por terem-me acolhido como a uma filha.

A Deus, por ter-me dado a bênção de ter sempre o que agradecer.

BIOGRAFIA

TELMA SUELY DA SILVA MORAIS, filha de Antônio da Silva Morais e Solange da Silva Morais (in memoriam), nasceu em Itabuna, Bahia, no dia 1º de julho de 1971.

Em março de 1994, iniciou o curso de Matemática na Universidade Federal de Viçosa, concluindo-o em dezembro de 2001.

Em março de 2007, ingressou no Programa de Pós-Graduação, em nível de mestrado, em Estatística Aplicada e Biometria, da Universidade Federal de Viçosa, submetendo-se à defesa da dissertação no dia 05 de dezembro de 2008.

SUMÁRIO

RESUMO

MORAIS, Telma Suely da Silva, M. Sc., Universidade Federal de Viçosa, dezembro de 2008. **Abordagem Bayesiana do modelo AR(1) para dados em painel: uma aplicação em dados temporais de microarray**. Orientador: Fabyano Fonseca e Silva. Co-Orientadores: Carlos Henrique Osório Silva, José Ivo Ribeiro Júnior e Paulo Roberto Cecon.

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 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 Generalizada foi usada para representar 3 diferentes prioris: simétrica, assimétrica e constante. 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.

ABSTRACT

MORAIS, Telma Suely da Silva, M. Sc., Universidade Federal de Viçosa, December of 2008. **Bayesian approach of AR(1) panel data model: application in microarray time series data**. Adviser: Fabyano Fonseca e Silva. Co-Advisers: Carlos Henrique Osório Silva, José Ivo Ribeiro Júnior and Paulo Roberto Cecon.

We considered a Bayesian analysis of first order autoregressive, AR(1), panel data model, using exact likelihood function, comparative analysis of prior distributions 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 flat 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.

1. INTRODUÇÃO

Uma Série Temporal é uma coleção de observações tomadas sequencialmente ao longo do tempo, e sua mais importante característica é que as observações vizinhas são consideradas dependentes, portanto, tem-se o interesse em aplicar técnicas estatísticas para analisar e modelar esta dependência.

A grande área da estatística denominada de Séries Temporais engloba uma gama enorme de modelos, cada um apresentando características próprias, que os diferenciam em relação às suas complexidades de análises. Dentre estes, um dos mais utilizados é o modelo auto-regressivo, denotado por AR(p), em que p é o número de parâmetros considerados além do intercepto. Sua generalização mais simples é dado pelo auto-regressivo de primeira ordem, AR (1). A grande vantagem do modelo em questão é que este se aplica a diversas situações práticas e geralmente apresenta boa qualidade de ajuste quando comparado com modelos mais complexos.

Uma técnica estatística com grande aplicação nas áreas de Econometria, Ciências Sociais e, mais recentemente, em Ciências Biológicas, é caracterizada pela modelagem simultânea de várias Séries Temporais. Esta técnica recebe o nome de análise de dados em painel, e embora apresente maior complexidade em relação aos estudos individuais de cada série, seu processo de estimação é geralmente mais eficiente, uma vez que utiliza informações de todo o conjunto de dados para estimar parâmetros individuais de cada série.

Em recentes estudos a metodologia Bayesiana foi utilizada com sucesso na análise de Séries Temporais (SILVA, 2006) e no ajuste de modelos de regressão não

linear (SAVIAN, 2008), pois segundo os autores, por considerar todos os parâmetros como variáveis aleatórias, a metodologia Bayesiana reduziu substancialmente o número de estimativas viesadas. Além disso, os autores relatam também que a Inferência Bayesiana geralmente demanda um número menor de observações, pois os conceitos probabilísticos envolvidos diminuem a dependência do ajuste do modelo em relação ao número de dados utilizados, uma vez que o conceito de graus de liberdade não é considerado.

Em termos teóricos, uma suposta vantagem da Inferência Bayesiana em relação à Inferência Clássica, é a utilização da distribuição a priori, a qual permite que o pesquisador incorpore seu conhecimento prévio a respeito da distribuição do parâmetro a ser estimado. Sendo assim, estudos foram conduzidos (SILVA, 2006; SAVIAN, 2008) com o intuito de selecionar distribuições a priori por meio do cálculo de estatísticas descriminantes como o Fator de Bayes e Pseudo-Fator de Bayes.

Apesar de apresentar as vantagens relatadas anteriormente, muitas vezes para se utilizar a Inferência Bayesiana, dependendo da complexidade do modelo adotado, conhecimentos avançados a respeito de Distribuições de Probabilidade são requeridos, e este fato pode prejudicar a utilização desta metodologia, principalmente por parte de profissionais que atuam em áreas aplicadas. Outro fato relevante é que ao se trabalhar com comparação de distribuições a priori, para cada distribuição de probabilidade considerada, deve-se realizar todo o procedimento teórico relacionado com a aplicação do Teorema de Bayes, isto é: Posteriori $\propto$ Priori x Verossimilhança. Além disso, também se deve levar em consideração os aspectos computacionais, que dependendo da complexidade das distribuições adotadas, geralmente apresentam problemas associados com a convergência dos algoritmos utilizados no processo de estimação. Portanto, estes fatores podem ocasionar aumento na demanda de tempo e recursos para a conclusão de pesquisas na área Estatística Bayesiana.

Uma possível solução para este problema é utilizar uma mesma distribuição de probabilidade a priori, e adotar diferentes valores para os seus parâmetros, denominados hiperparâmetros, de forma que esta passe a representar diferentes formas, e assim caracterizar outras distribuições de probabilidade. Uma distribuição que pode ser utilizada para este objetivo é a Beta, pois esta apresenta uma função densidade de probabilidade (f.d.p.) relativamente simples e flexível (SILVA, 2000).

Pesquisas relacionadas com a aplicação de métodos Bayesianos em estudos de Séries Temporais (SILVA, 2006) e dados longitudinais (SAVIAN, 2008), 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, pode-se citar a análise de dados de Expressão Gênica periodicamente identificados ao longo do tempo, os quais são denominados Microarray Time Series (MTS) (MUKHOPADHYAY; CHATTERJEE, 2007).

Com relação aos MTS, Fujita et al. (2007) compararam a eficiência de diferentes modelos fundamentados na teoria de vetores auto-regressivos (VAR – Vector Autoregressive), os quais estão associados com a análise multivariada de Séries Temporais. Estes autores utilizaram estes modelos para descrever o comportamento de índices de expressão gênica ao longo de diferentes horas. Silva (2006) relata que a análise de modelos auto-regressivos para dados em painel é uma técnica mais simples, porém não menos eficiente que a técnica VAR, o que possibilita sua aplicação aos dados MTS, fato este do qual até o presente momento ainda não se tem relato na literatura especializada.

O objetivo deste trabalho foi realizar a análise Bayesiana do modelo auto-regressivo de primeira ordem para dados em painel e comparar distribuições a priori caracterizadas por diferentes distribuições Beta (simétrica, assimétrica e constante). Objetivou-se também validar a metodologia empregada via simulação de dados e aplicá-la em um conjunto de dados MTS.

2. REVISÃO DE LITERATURA

2.1. Séries temporais e dados em painel

No final dos anos 80 desenvolveu-se um interesse crescente pela utilização de dados longitudinais nas diversas áreas do conhecimento, e este fenômeno foi impulsionado, sobretudo pelas possibilidades de se estimar modelos de comportamento individual com dados agregados de várias Séries Temporais (ARELLANO; BOVER, 1990).

Sob o ponto de vista estatístico, a grande motivação para o uso de dados em painel surge em estudos onde se observam problemas de estimação relacionados a muitas séries com pouco número de observações, pois a integração dos dados de todas as séries fornece um aumento considerável na quantidade de informação disponível para a análise (KITTEL, 1999).

Segundo Baltagi (2000), a maior quantidade de informação disponibilizada pela estrutura em painel aumenta a eficiência da estimação, uma vez que permite identificar e medir efeitos que não serão pura e simplesmente detectáveis em estudos exclusivamente seccionais ou temporais.

Um exemplo prático de análise de dados em painel á apresentado por Hirano (2002). De acordo com este autor, a estrutura em painel fornece meios de se realizar simultaneamente uma análise "cross-section" e uma análise temporal. O mesmo considera dados referentes a Pesquisas Industriais Anuais (PIA), e afirma que é possível fazer análise para um determinado ano usando várias firmas, isto é, usando dados em

“cross-section”, e também para uma firma em vários anos, cuja estrutura se caracteriza como dados temporais, porém uma forma mais complexa, e mais informativa, é realizar a análise para várias firmas em vários anos conjuntamente, fato que define a estrutura em painel.

Segundo Frees (2004), os dados em painel proporcionam uma maior quantidade de informação, maior número de graus de liberdade e maior eficiência na estimação. Adicionalmente, os estudos com amostras longitudinais possibilitam uma análise mais eficiente do ajustamento, pois os estudos seccionais, ao não contemplarem a variabilidade longitudinal, transmitem uma falsa idéia de estabilidade. Assim, a utilização de dados em painel permite conjugar a diversidade de comportamentos individuais, com a existência de um padrão, ou seja, permite tipificar as respostas de diferentes indivíduos.

Em relação a algumas desvantagens de se utilizar uma análise de dados em painel, Hsiao e Sun (2000) relatam que se considerarmos uma população como um conjunto de decisões que se refletem em diferentes histórias individuais, estas terão que ser representadas como variáveis aleatórias específicas a cada indivíduo e que certamente estarão correlacionadas não apenas com a variável dependente, mas também com as características que envolvem o indivíduo, fato este que pode gerar dificuldades para a especificação dos modelos e para a estimação de parâmetros. Para evitar este problema, os autores recomendam que é necessário optar por um conjunto de indivíduos homogêneos, que possam realmente ser identificados como elementos de uma mesma população.

Com relação à formação de grupos de indivíduos homogêneos, Silva (2006) relata que ao estudar dados de mérito genético de vários touros ao longo do tempo por meio de um modelo AR (2) para dados em painel, o agrupamento de touros de acordo com a acurácia de seus valores genéticos estimados possibilitou melhorar significativamente a qualidade de previsão de valores futuros.

2.2. Inferência Bayesiana

2.2.1. Definições

Fazer inferências é uma das principais finalidades da estatística. Na abordagem clássica, os parâmetros desconhecidos são considerados fixos e toda a análise é baseada

nas informações contidas na amostra dos dados. Segundo Paulino et al. (2003), esta abordagem foi adotada de forma quase unânime pelos estatísticos durante a primeira metade do século XX. No entanto, a abordagem Bayesiana renasceu nos últimos anos, depois de longos períodos esquecida pelos pesquisadores.

A inferência Bayesiana trata o vetor de parâmetros desconhecidos como quantidades aleatórias e, consequentemente, permite obter algum conhecimento sobre esses antes que os dados tenham sido coletados, atribuindo assim distribuições de probabilidade. Elas podem ser obtidas através de análises anteriores, experiência do pesquisador na área em questão ou publicações sobre o assunto que se deseja tratar.

A inferência Bayesiana consiste de uma informação da distribuição a priori, dos dados amostrais e do cálculo da densidade a posteriori dos parâmetros. A informação da distribuição a priori é dada pela densidade de probabilidade $P(\theta)$, a qual expressa o conhecimento do pesquisador sobre os parâmetros a serem estimados. Quando em determinado estudo o pesquisador tem pouca ou nenhuma informação para incorporar a distribuição a priori, ou quando não há um modelo de distribuição que possa ser considerado como o mais adequado, resta ainda a alternativa de se utilizar uma não informativa, por exemplo, distribuição a priori de Jeffreys (JEFFREYS, 1961). Os dados $Y=\{y_1, y_2, ..., y_n\}$, representados por uma amostra aleatória de uma população com densidade f, são utilizados na análise Bayesiana na função de verossimilhança $L(y_1, ..., y_n \mid \theta)$, que é a densidade conjunta destes dados.

Portanto, a partir do momento que se opta por uma distribuição a priori, seja ela informativa ou não, e obtém-se a função de verossimilhança, é possível, por meio do Teorema de Bayes representado pela expressão (1), obter a função densidade a posteriori de θ, de forma que qualquer inferência sobre θ é realizada a partir desta distribuição,

$$P(\theta \mid Y_n) = \frac{L(Y_n \mid \theta)P(\theta)}{\int L(Y_n \mid \theta)P(\theta)d\theta}, \qquad (1)$$

sendo $Y_n = \{y_1, y_2, ..., y_n\}$. O denominador, chamado de constante de integração, não depende de θ, portanto temos:

$$P(\theta \mid Y) \propto L(Y_n \mid \theta)P(\theta), \qquad (2)$$

ou seja: Posteriori $\propto$ Verossimilhança x Priori, em que $\propto$ representa proporcionalidade.)

Com já relatado anteriormente, a distribuição a posteriori de um parâmetro contém toda a informação probabilística a respeito do mesmo. Segundo Rosa (1998), para se inferir em relação a qualquer elemento de $\boldsymbol{\theta}$, a distribuição a posteriori conjunta dos parâmetros, P($\boldsymbol{\theta}$ |Y), deve ser integrada em relação a todos os outros elementos que a constituem. Assim, se o interesse do pesquisador se concentra sobre determinado conjunto de $\boldsymbol{\theta}$, por exemplo, θ_1, tem-se a necessidade da obtenção da distribuição p(θ_1 |Y), denominada de marginal, a qual é dada por:

$$P(\theta_1 \mid Y) = \int_{\boldsymbol{\theta} \neq \theta_1} P(\boldsymbol{\theta} \mid Y) d\boldsymbol{\theta}_{\boldsymbol{\theta} \neq \theta_1} \tag{3}$$

A integração da distribuição conjunta a posteriori para a obtenção das marginais geralmente não é analítica, necessitando de algoritmos iterativos especializados como o Gibbs Sampler e o Metropolis-Hastings, os quais são denominados de algoritmos MCMC (Markov Chain-Monte Carlo). Portanto, para a utilização desses algoritmos, é necessário que se obtenha a partir da distribuição a posteriori um conjunto de distribuições chamadas de distribuições condicionais completas.

2.2.2. Métodos de Monte Carlo via Cadeias de Markov (MCMC)

Uma cadeia de Markov é um processo estocástico iterativo no qual se assume que o próximo estado da cadeia, θ^{t+1}, depende somente do estado atual θ^t e dos dados, e não da história passada da cadeia .Existe uma dependência do estado inicial da cadeia, θ^0, que é esquecido após um período de aquecimento denominado burn-in, o qual é geralmente descartado da análise. Um outro detalhe também utilizado é o espaçamento entre as iterações sucessivas denominado thin, usado para eliminar a autocorrelação entre as iterações consecutivas.

O objetivo dos métodos MCMC é obter uma amostra das distribuições marginais a posteriori dos parâmetros de interesse por meio de um processo iterativo utilizando as distribuições condicionais completas de cada parâmetro. Por sua vez, esses valores gerados são considerados amostras aleatórias de uma determinada distribuição de probabilidade, caracterizando assim o método de simulação Monte Carlo.

O algoritmo de Metropolis-Hastings permite gerar uma amostra da distribuição conjunta a posteriori $\pi(\theta_1, \theta_2, ..., \theta_k \mid x)$, a partir das distribuições condicionais completas, que podem possuir forma fechada ou não (METROPOLIS et al., 1953;

HASTINGS, 1970). Esses autores usam a idéia de que um valor é gerado de uma distribuição auxiliar ou candidata e este é aceito com uma dada probabilidade. Um caso particular é chamado de algoritmo de Metropolis-Hastings e considera apenas propostas simétricas, ou seja, $q(\theta^{'} \mid \theta_1) = q(\theta_1 \mid \theta^{'})$ para todos os valores de θ_1 e $\theta^{'}$. Neste caso a probabilidade de aceitação se reduz para $\alpha(\theta_1, \theta^{'}) = \min(1, \frac{p(\theta_1^{'} \mid \theta_2, ..., \theta_d)}{p(\theta_1 \mid \theta_2, ..., \theta_d)})$. Uma apresentação descritiva deste algoritmo é mostrada no Quadro 1.

Quadro 1 – Descrição do algoritmo de Metropolis-Hastings

I - Inicialize o contador de iterações $t = 0$ e especifique valores iniciais $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})$, onde d representa o número de parâmetros;

II - Gere um novo valor $\theta^{'}$ da distribuição proposta $q(. \mid \theta_1)$;

III - Calcule a probabilidade de aceitação $\alpha(\theta, \theta^{'})$ e gere $u \sim U(0,1)$;

$$\alpha(\theta_1, \theta^{'}) = \min(1, \frac{\pi(\theta^{'} \mid \theta_2, ..., \theta_d) q(\theta_1 \mid \theta^{'})}{\pi(\theta \mid \theta_2, ..., \theta_d) q(\theta^{'} \mid \theta_1)})$$

IV - Se $u \leq \alpha$ então aceite o novo valor e faça $\theta^{(t+1)} = \theta^{'}$, caso contrário rejeite e faça $\theta^{(t+1)} = \theta$ **;**

V - Incremente o contador de t para $t+1$ e volte ao passo **II**.

O amostrador de Gibbs é um caso especial do Metropolis-Hastings, que permite gerar uma amostra da distribuição conjunta a posteriori $\pi(\theta_1, \theta_2, ..., \theta_k \mid x)$, desde que as distribuições condicionais completas possuam forma fechada, no sentido que seja fácil amostrar de seus elementos (GELFAND, 2000). Uma apresentação descritiva do algoritmo em questão é mostrada no Quadro 2.

Quadro 2 – Descrição do algoritmo Gibbs Sampler (Amostrador de Gibbs)

I - Inicialize o contador de iterações da cadeia $t = 0$;

II - Especifique valores iniciais $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})$;

III - Obtenha um novo valor de $\theta^{(t)}$ a partir de $\theta^{(t-1)}$ através da geração sucessiva de valores

$$\theta_1^t \sim \pi(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_d^{(t-1)})$$
$$\theta_2^t \sim \pi(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_d^{(t-1)})$$
$$...$$
$$\theta_d^t \sim \pi(\theta_d \mid \theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{d-1}^{(t)})$$

IV - Incremente o contador de t para $t+1$ e retorne ao passo **II** até obter convergência.

Os algoritmos MCMC são processos interativos, portanto surgem questões referentes à avaliação de suas convergências. Na literatura são apresentados vários métodos necessários para a realização desta avaliação, e dentre estes se destacam Geweke (1992), Gelman e Rubin (1992), Raftery e Lewis (1992) e Heidelberger e Welch (1993). Uma maneira prática de aplicar todos estes métodos é por meio do pacote BOA (SMITH, 2007) do software livre R (R DEVELOPMENT CORE TEAM, 2008).

2.2.3. Comparação de Prioris

Recentemente, pesquisadores que utilizaram a inferência Bayesiana (SILVA, 2006; REIS, 2008; SAVIAN, 2008) optaram por realizar comparações de distribuições a priori a fim de indicar aquela que é mais plausível para representar o conhecimento prévio a respeito dos parâmetros a serem estimados. Em se tratando de modelos auto-regressivos, vários autores (ZELLNER, 1996; VERMAAK et al., 2000; SÁFADI; MORETTIN, 2003) utilizaram uma distribuição normal como priori para os parâmetros de auto-regressão, respeitando é claro, o espaço paramétrico relacionado com a ordem do modelo. Porém, segundo vários autores, dentre eles Barreto e Andrade (2004), Falk e

Roy (2005), Zhou e Roy (2006) e Silva (2006), a distribuição t-Student foi empregada com sucesso na descrição de coeficientes de modelos auto-regressivos complexos, e segundo estes autores, se deve ao fato desta distribuição apresentar uma cauda mais pesada, a qual lhe confere uma maior capacidade em expressar o desconhecimento à priori em relação aos parâmetros de interesse.

Ghosh e Heo (2003) utilizaram simulação de dados, com ênfase na probabilidade de cobertura dos intervalos de credibilidade, para comparar vários tipos de distribuições a priori para um modelo auto-regressivo de primeira ordem, AR (1), e concluíram que certas distribuições a priori informativas apresentaram resultados melhores que a distribuição a priori não informativa de Jeffreys. Ni e Sun (2003), estudando modelos auto-regressivos multivariados, compararam diferentes distribuições a priori para o vetor de coeficientes. Esta comparação foi realizada com base na estabilização dos momentos da distribuição a posteriori, e indicou que a distribuição a priori de referência proposta por Yang e Berger (1994) apresentou melhor resultado.

Ao se utilizar diferentes distribuições a priori para o mesmo parâmetro de um modelo, tem-se, sob o ponto de vista Bayesiano, diferentes modelos, que por sua vez podem ser comparados. Esta comparação muitas vezes é realizada por meio do fator de Bayes (KASS; RAFTERY, 1995), o qual de forma geral, corresponde a uma razão de distribuições a posteriori, seguindo um processo análogo ao de razão de verossimilhanças sob o enfoque Freqüentista.

De acordo com Kass e Raftery (1995), o fator de Bayes (FB) é definido por: $FB_{ij} = \frac{P(\boldsymbol{Y}|M_i)}{P(\boldsymbol{Y}|M_j)}$, em que: $P(\boldsymbol{Y}|M_p)$ é o fator de normalização, e corresponde ao denominador da expressão (1). Esta quantidade recebe este nome, pois é ela que garante a característica de densidade de probabilidade à distribuição a posteriori, impedindo que esta distribuição seja considerada imprópria. Na literatura (KASS; RAFTERY, 1995; RAFTERY, 1996) esta quantidade também é geralmente denominada de Verossimilhança Marginal, e pode ser obtida da seguinte forma:

$$P(\boldsymbol{Y}|M_p) = \int_{\boldsymbol{\theta}} L(\boldsymbol{Y}|\boldsymbol{\theta}, M_p) P(\boldsymbol{\theta}|M_p) d\boldsymbol{\theta} \qquad (4)$$

em que: $L(\boldsymbol{Y}|\boldsymbol{\theta}, M_p)$ e $P(\boldsymbol{\theta}|M_p)$ são, respectivamente, a função de verossimilhança e a distribuição a priori correspondentes ao modelo M_p.

A resolução da integral apresentada na equação (4) geralmente não é analítica, mas segundo Raftery (1996) uma possível solução é considerar valores de $\boldsymbol{\theta}^{(k)}$ gerados via algoritmos MCMC, pois assim é possível obter uma estimativa de $P(\boldsymbol{Y}|M_p)$, a qual é obtida por: $\hat{P}(\boldsymbol{Y}|M_p) = \frac{1}{Q}\sum_{q=1}^{Q} L(\boldsymbol{Y}|\boldsymbol{\theta}^{(q)}, M_p)$, em que o índice q representa cada iteração dos algoritmos MCMC, sendo q = 1,2,...,Q.

Em relação à interpretação do FB, pode-se dizer que esta é simples e direta, pois se:

FB > 1 tem-se a indicação que o modelo disposto no numerador é melhor, caso contrário, se FB < 1, o modelo do denominador é o preferido. Se FB=1, a qualidade dos dois modelos é a mesma.

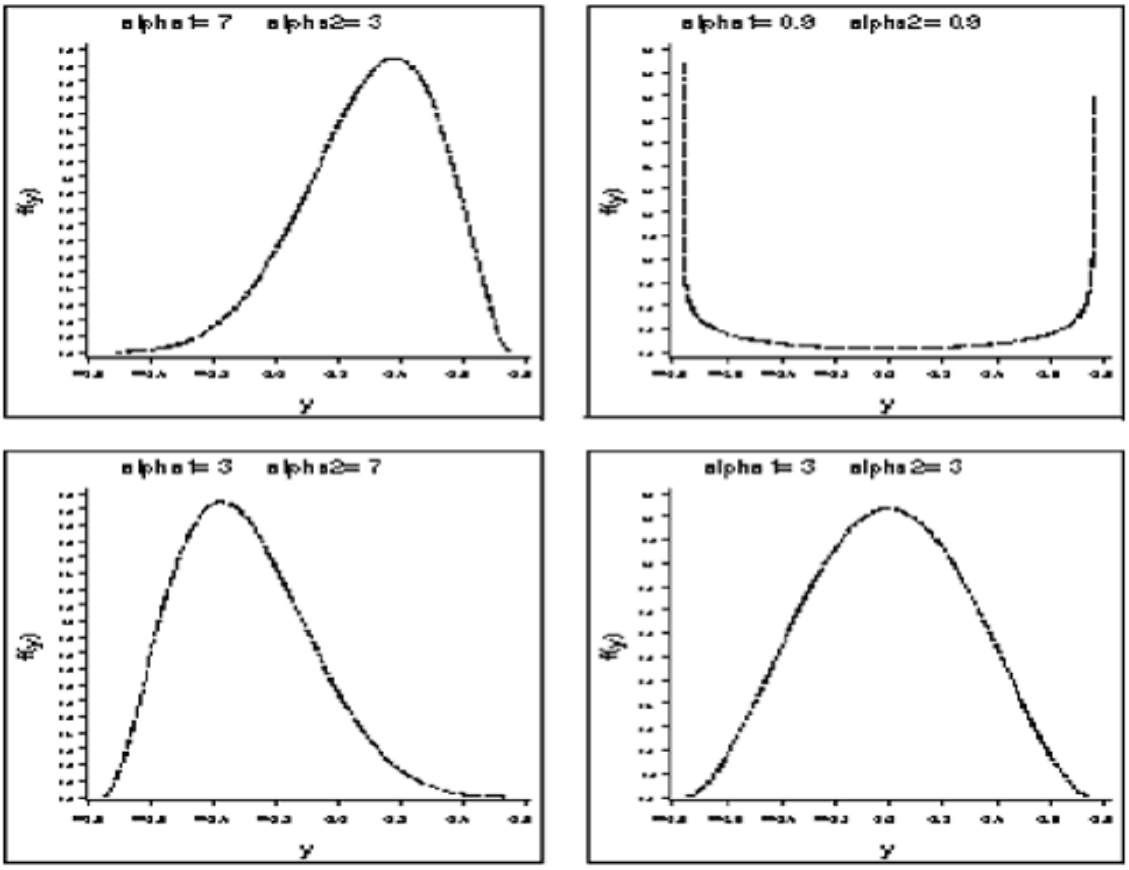
Além do Fator de Bayes possibilitar a comparação entre modelos caracterizados por diferentes distribuições a priori, de acordo com Paulino et al. (2003), o mesmo também pode ser usado para comparar modelos identificados pela mesma distribuição a priori, porém com hiperparâmetros diferentes. Neste contexto, os hiperparâmetros são representados pelos valores fixos assumidos para os parâmetros da distribuição a priori.

2.3. Características da Distribuição a priori Beta

A respeito da Inferência Bayesiana, Turkman (2008) aponta alguns argumentos contrários a sua aplicação, e dentre estes se destaca a dificuldade na formulação de prioris quanto aos aspectos teóricos, ou seja, muitas distribuições de probabilidade, ao serem usadas como distribuição a priori, podem gerar distribuições a posteriori com alta complexidade, proporcionando problemas na obtenção das distribuições condicionais completas e na implementação dos algoritmos MCMC. Assim, estes problemas podem determinar um aumento no tempo de conclusão de pesquisas e nos custos referentes a elas, uma vez que serão exigidos recursos computacionais de alta qualidade para suportar complexas implementações dos algoritmos MCMC.

Em relação a comparação Bayesiana de modelos, ao se utilizar diferentes distribuições a priori, são requeridos diferentes desenvolvimentos do teorema de Bayes, Posteriori $\propto$ Priori x Verossimilhança, isto é, para cada distribuição a priori considerada é necessário obter uma distribuição a posteriori. Uma forma viável de facilitar este processo é utilizar uma mesma distribuição de probabilidade a priori, e adotar diferentes

valores para os seus parâmetros, os quais são definidos como hiperparâmetros, de forma que esta passe a representar diferentes formas, e assim caracterizar outras distribuições de probabilidade. Dentre as distribuições que apresentam esta qualidade, destaca-se a Beta, pois sua função densidade de probabilidade (f.d.p.) é relativamente simples e dispõe de grande flexibilidade, Casella & Berger (1994) citados por (Silva, 2000). Em outras palavras, a distribuição Beta pode adquirir facilmente a forma de distribuições simétricas, assimétricas e constantes mediante simples modificações nos valores de seus parâmetros. Exemplos de gráficos das f.d.p que ilustram a flexibilidade relatada são apresentados na Figura 1.



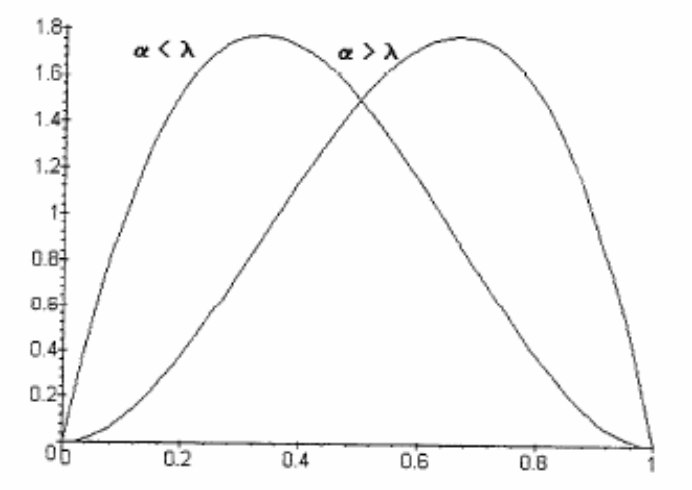
Fonte: SILVA, 2000.

Figura 1 – Diferentes formas assumidas pela distribuição Beta de acordo com alterações nos valores de seus parâmetros.

Segundo Bearzotti (1998), originalmente a distribuição beta foi concedida para Y contida no intervalo entre zero e um. Sua f.d.p. utiliza 2 parâmetros α e λ, e é dada por:

$$f(y;\alpha,\lambda)=\frac{1}{B(\alpha,\lambda)}y^{\alpha-1}(1-y)^{\lambda-1}I_{[0,1]}(y), \text{ em que } B(\alpha,\lambda)=\frac{\Gamma(\alpha)\Gamma(\lambda)}{\Gamma(\alpha+\lambda)} \qquad (5)$$

com as restrições de que α e λ sejam maiores do que zero. Se α =λ=1, então a distribuição é chamada retangular, pois corresponde a uma linha paralela ao eixo x em f(y)=1, e corresponde a distribuição Uniforme. Se α>0 e λ<0, então f(y) é crescente; se α<0 e λ>0, então é decrescente, e se os parâmetros forem negativos ela apresentará uma aparência de "U". Sua maior utilização, entretanto, corresponde aos casos de α e λ maiores que zero, dando uma aparência como a da Figura 2.



Fonte: BEARZOTTI, 1998.

Figura 2 – Diferentes formas assumidas pela distribuição Beta de acordo com alterações nos valores de seus parâmetros.

Aplicando a definição da função Beta no cálculo das esperanças de Y e Y², facilmente verifica-se que:

$$E(Y)=\frac{\alpha}{\alpha+\lambda},\ E(Y^2)=\frac{\alpha(\alpha+1)}{(\alpha+\lambda)(\alpha+\lambda+1)} \text{ e } V(Y)=\frac{\alpha\lambda}{(\alpha+\lambda)^2(\alpha+\lambda+1)}.$$

É possível aumentar as aplicações da distribuição Beta por meio de uma generalização para casos em que a variável aleatória é definida em um intervalo [a,b] qualquer, não necessariamente entre 0 e 1. Nessa generalização, tem-se que a f.d.p. passa a ser dada por (SCOLFORO,1995):

$$f(y;\alpha,\lambda,a,b)=\frac{1}{B(\alpha,\lambda)}\frac{1}{(b-a)^{\alpha+\lambda-1}}(y-a)^{\alpha-1}(b-y)^{\lambda-1}I_{[a,b]}(y) \qquad (6)$$

Quanto à distribuição de probabilidade Beta Generalizada (6), esta foi utilizada com sucesso por Morais et al. (2008) para representar a distribuição a priori do coeficiente do modelo auto-regressivo de primeira ordem. Neste caso, os valores de a e b foram, respectivamente, -1 e 1, respeitando assim o espaço paramétrico deste coeficiente.

2.4. Dados MTS (Microarray Time Series)

Bioinformatica é a área que estuda o projeto e implementação de novos métodos computacionais para auxiliar pesquisadores de Biologia Molecular, Bioquímica, e outras áreas na análise de grandes quantidades de dados biológicos. Os avanços tecnológicos e da pesquisa em áreas como genômica, transcriptoma e proteômica têm produzido massas enormes de dados que precisam de uma análise mais aprofundada para sua compreensão. Assim surge a necessidade da utilização de técnicas estatísticas, matemáticas e procedimentos computacionais que auxiliem os pesquisadores a extrair destes dados informações importantes a respeito de processos biológicos (FUJITA, 2007).

Alguns dos problemas mais importantes que se estuda nesta área estão ligados à compreensão do funcionamento do sistema celular. Sabe-se da importância dos genes neste sistema, mas ainda é tema de pesquisa intensa entender como seus produtos interagem para desencadear os processos celulares, isto é, como os genes são ativados e expressos ao longo do tempo. A expressão gênica é regulada por complexas redes de interação entre DNA, RNA, proteínas e outras moléculas pequenas, as quais se interagem com o tempo, sendo muito difícil uma compreensão intuitiva do seu dinamismo (JONG, 2002). De acordo com Faceli et al. (2005) e Fujita (2007), atualmente a modelagem estatística deste fenômeno apresenta grande relevância, uma vez que permite predizer e simular o comportamento da expressão gênica sem ambigüidades e de forma sistemática.

Segundo Fujita et al. (2007), a modelagem em questão é feita a partir de dados provenientes de técnicas de expressão, dentre as quais se destaca o microarray. O microarray é uma técnica desenvolvida recentemente, que tem se mostrado bastante útil para a quantificação simultânea dos níveis de expressão gênica, inclusive ao longo do tempo, de milhares de genes. Quando estes dados temporais são avaliados, os mesmos podem ser denominados de Microarray Time Series (MTS) (MUKHOPADHYAY; CHATTERJEE, 2007).

Segundo Yamaguchi et al. (2007) os microarrays permitem medir simultaneamente a resposta ou o nível de expressão de um elevado número de genes em determinadas condições experimentais, as quais podem corresponder a diferentes instantes de tempo. A modelagem deste tipo de dados pode ser especialmente útil para a identificação de conjuntos de genes com comportamentos semelhantes em determinados instantes temporais. Dessa forma, é possível pensar na utilização de modelos de séries temporais para dados em painel, pois estes consideram a modelagem simultânea de várias séries, que nesta aplicação correspondem às medidas de expressão de cada gene em diferentes tempos.

Ainda em relação ao segundo ponto do parágrafo anterior, Faceli et al. (2005) relatam que os dados de expressão gênica são geralmente representados em uma matriz, com as linhas representando os genes e as colunas representando as amostras (diferentes tecidos, estágios de desenvolvimento, tratamentos, instantes de tempo, dentre outros). Essa matriz é chamada matriz de expressão e devido ao alto custo dos experimentos, esta é, geralmente, constituída de um grande número de linhas (genes) e poucas colunas (amostras).

De acordo com exposto, se faz necessário o desenvolvimento e a comparação de técnicas estatísticas de análise de dados capazes de suprir estas falhas. Seguindo esta linha, Fujita (2007) comparou as seguintes metodologias fundamentadas na modelagem de vetores auto-regressivos (VAR): SVR (Support Vector Regression), DVAR (Dynamic Vector Auto Regressive Model) e SVAR (Sparse Vector Auto Regressive Model). No estudo em questão, tais metodologias foram utilizadas para descrever o comportamento da expressão de vários genes ao longo de diferentes horas sobre células HeLa (células humanas epiteliais provenientes da fase final de crescimento).

Com relação aos modelos VAR, Silva (2006) descreve que estes constituem alternativas viáveis, porém bem mais complexas, aos modelos auto-regressivos para dados em painel. Assim, subtende-se que estes últimos podem vir a caracterizar uma classe de modelos mais simples a serem utilizados na análise de dados MTS.

3. MATERIAL E MÉTODOS

3.1. Modelo auto-regressivo para dados em painel

O modelo auto-regressivo de ordem p, AR(p), em que p é o número de parâmetros considerados, para uma única série temporal é o seguinte (MORETTIN; TOLOI, 2004):

$$Z_j = \phi_1 Z_{j-1} + \phi_2 Z_{j-2} + ... + \phi_p Z_{j-p} + e_j = \sum_{k=1}^{p} \phi_k Z_{j-k} + e_j \quad j=1,2,3,...n\,, \tag{7}$$

em que:

Z_j é o valor da série numérica dada por Z_{j-1}, Z_{j-2}, ... , Z_{j-p};

ϕ_1, ϕ_2,..., ϕ_p são os parâmetros do modelo, denominados de parâmetros de auto-regressão;

e_j é o erro aleatório não observável, associado à observação Z_j, também denominado de ruído branco, assumido como $e_j \overset{iid}{\sim} N(0, \sigma_e^2)$.

Na presença de dados em estrutura de painel, o modelo AR(p) indicado em (7) continua apresentando as mesmas características, porém a ele deve ser acrescentado um índice i referente ao indivíduo (LIU; TIAO, 1980), caracterizando assim a seguinte expressão:

$$Z_{ij} = \phi_{i1} Z_{i(j-1)} + \phi_{i2} Z_{i(j-2)} + ... + \phi_{mp} Z_{m(j-p)} + e_{ij} = \sum_{k=1}^{p} \phi_{ik} Z_{i(j-k)} + e_{ij}. \tag{8}$$

em que: i=1,2,...m; k=1,2,...,p e j=1,2,...,n_i. Assume também, que: $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$.

De acordo com esta notação, têm-se m indivíduos, com n_i observações longitudinais cada um, indicando que cada indivíduo i pode apresentar um número diferente de observações. Nota-se também que o modelo contempla p parâmetros por indivíduo.

O modelo apresentado em (8) contempla a forma completa dos modelos auto-regressivos para dados em painel, porém sua caracterização mais simples é dada quando p=1, sendo este o modelo AR (1) para dados em painel:

$$Z_{ij} = \phi_i Z_{i(j-1)} + e_{ij},\ i = 1,2,3,...,m; j = 1,2,3,...,n_i. \quad (9)$$

No modelo apresentado em (9), assumiu-se que $e_{ij} \sim N(0,\sigma^2)$, ou seja, que $Z_{ij} \sim N(\phi_i Z_{i(j-1)}, \sigma^2)$.

3.2. Função de verossimilhança

No que se segue, será considerado $n_1 = n_2 = ...= n_m = n$), ou seja, o mesmo número de observações longitudinais para cada indivíduo i, i=1, 2,...,m. A partir da distribuição de Z_{ij} é possível construir a função de verossimilhança, dada por $P(Z_i \mid \phi_i, \sigma^2)$, a qual representa a distribuição conjunta dos dados de cada série. No presente trabalho adotou-se a função de verossimilhança exata (10), uma vez que esta considera um termo referente a primeira observação de cada série, $P(Z_{i1} \mid \phi_i, \sigma^2)$, e um outro termo denominado de função de verossimilhança aproximada ou condicional, dada por $P(Z_i^* \mid \phi_i, \sigma^2)$. Neste último, o índice j=1 não é considerado devido a defasagem observada no índice j do termo $Z_{i(j-1)}$ em (9), uma vez que não é possível identificar a observação $Z_{i(1-1)} = Z_{i0}$. Dessa forma, tem-se:

$$P(Z_i \mid \phi_i, \sigma^2) = P(Z_{i1} \mid \phi_i, \sigma^2) \times P(Z_i^* \mid \phi_i, \sigma^2) \quad (10)$$

$$P(Z_{i1} \mid \phi_i, \sigma^2) \propto \left(\frac{1-\phi_i^2}{\sigma^2}\right)^{\frac{1}{2}} .\exp\left\{-\frac{1}{2\sigma^2}\left(Z_{i1}^2\left(1-\phi_i^2\right)\right)\right\}$$ (Morettin e Toloi, 2004)

$$P(Z_i^* \mid \phi_i, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n-1}{2}} .\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=2}^{n} Z_{ij}^2 - 2\phi_i \sum_{j=2}^{n} Z_{ij} Z_{i(j-1)} + \phi_i^2 \sum_{j=2}^{n} Z_{i(j-1)}^2\right\}$$ (Liu e Tiao, 1980)

Denominando: $h_i = \sum_{j=2}^{n-1} Z_{ij}^2$ $\quad \hat{\phi}_i = \dfrac{\sum_{j=2}^{n} Z_{ij} Z_{i(j-1)}}{\sum_{j=2}^{n-1} Z_{ij}^2}$ $\quad S_i^2 = \sum_{j=1}^{n} Z_{ij}^2 - h_i \hat{\phi}_i^{\,2}$, tem-se que:

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

$$P(\mathbf{Z} \mid \boldsymbol{\phi}, \sigma^2) = \prod_{i=1}^{m} P(Z_i \mid \phi_i, \sigma^2) \qquad (11)$$

As demonstrações de como obter a expressão (11) são apresentadas no Apêndice 1.

3.3. Distribuições a Priori

3.3.1. Parâmetros da distribuição Beta Generalizada

Optou-se por utilizar a distribuição Beta Generalizada no intervalo [-1,1] como priori, de forma a assumir diferentes valores para os seus parâmetros α e λ. Este procedimento permite que uma mesma distribuição, no caso a Beta Generalizada, seja utilizada para representar outras distribuições com diferentes formas. Foram utilizados valores para α e λ que representassem as seguintes classes de distribuições: simétrica, assimétrica e constante.

Os valores de α e λ que caracterizam cada classe foram obtidos mediante sucessivas tentativas a fim de deixar a Beta com a forma de distribuições representativas de cada classe. Neste contexto, utilizaram-se as seguintes distribuições: normal truncada no intervalo [-1,1] para a simétrica; Gumbel truncada [-1,1] para a assimétrica e Uniforme [-1,1] para a constante. Os valores de α e λ foram escolhidos em função do bom ajuste da Beta Generalizada em relação a cada uma destas três distribuições apresentadas. A Figura 3 elucida este procedimento e mostra os respectivos parâmetros utilizados.

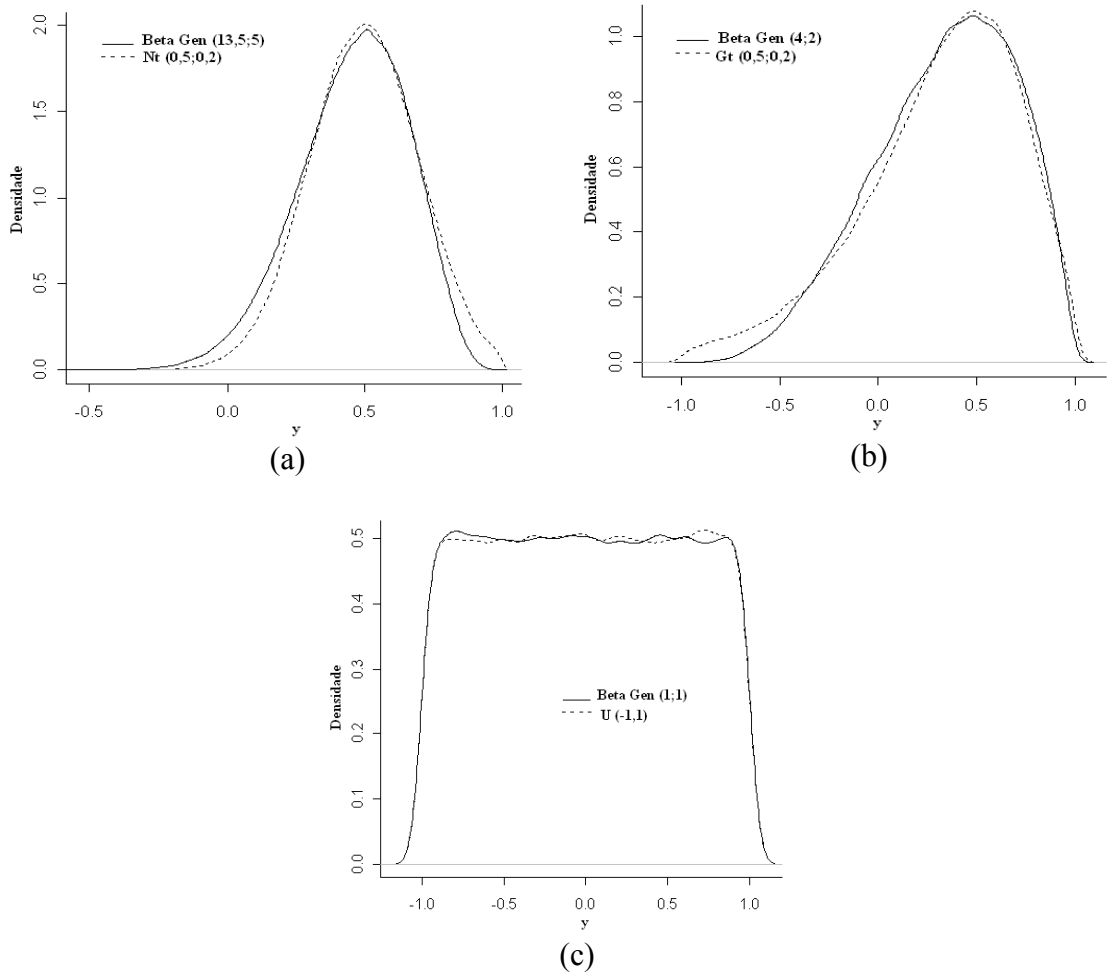


Figura 3 – Os gráficos a, b e c representam, respectivamente, as densidades da distribuição Beta Generalizada e seu ajuste às distribuições Normal, Gumbel e Uniforme.

3.3.2. Priori Beta Generalizada para ϕ_i

As distribuições Beta Generalizada mostradas na Figura 3 foram consideradas como distribuições a priori, assim a única diferença entre elas são os valores dos parâmetros α e λ. A expressão que representa a priori para ϕ_i é a seguinte:

$P(\phi_i|\alpha,\lambda)=\frac{1}{\beta(\alpha,\lambda)}.\frac{1}{(1-(-1))^{\alpha+\lambda-1}}.(\phi_i-(-1))^{\alpha-1}.(1-\phi_i)^{\lambda-1}$. Como a primeira parte desta expressão não contém o parâmetro de interesse ϕ_i, a mesma pode ser desconsiderada, ou seja:

$$P(\phi_i|\alpha,\lambda) \propto (\phi_i+1)^{\alpha-1}(1-\phi_i)^{\lambda-1} \quad (12)$$

Esta distribuição enfatiza um dos objetivos do presente trabalho, que consiste na utilização de uma única distribuição para representar várias outras com diferentes formas, visando-se com isto uma simplificação na aplicação da metodologia Bayesiana.

3.3.3. Priori Gama-Inversa para σ^2

Além dos coeficientes ϕ_i's, a variância residual também deve ser estimada, uma vez que esta não é assumida como conhecida, o que a caracteriza também como um parâmetro, e por isso exige que se adote uma distribuição a priori para ela. Geralmente a Distribuição Gama-Inversa é adotada para este fim (SILVA, 2006; REIS et al., 2008; SAVIAN, 2008). Esta é dada por: $P(\sigma^2|c,d) \propto \left(\sigma^2\right)^{-(c+1)} \exp\left\{-\frac{d}{\sigma^2}\right\}$,em que c e d são os hiperparâmetros.

3.4. Distribuição Conjunta a Posteriori

Seguindo o teorema de Bayes, Posteriori $\propto$ Verossimilhança x Priori, temos então que a distribuição a posteriori conjunta será obtida pelo produto da função de verossimilhança (item 3.2) com as distribuições a priori (itens 3.2.2 e 3.2.3). Assim, obtém-se:

$$P(\phi_i,\sigma^2|Z_i,\alpha,\lambda,c,d) \propto P(Z_i \mid \phi_i,\sigma^2)\times P(\phi_i \mid \alpha,\lambda)\times P(\sigma^2|c,d)$$

$$\begin{aligned} P(\phi_i,\sigma^2|Z_i,\alpha,\lambda,c,d) \;\propto\; & (\sigma^2)^{-\frac{n}{2}}(1-\phi_i^2)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[h_i(\phi_i^2 - 2\phi_i\hat{\phi}_i + \hat{\phi}_i^2)\right]\right\}\times \\ & (\phi_i+1)^{\alpha-1}(1-\phi_i)^{\lambda-1}\times(\sigma^2)^{-(c+1)}\exp\left\{\frac{d}{\sigma^2}\right\} \end{aligned} \tag{13}$$

Para que os parâmetros ϕ_i e σ^2 possam ser estimados, torna-se necessário obter as suas distribuições marginais a posteriori, ou seja, a marginal de ϕ_i é obtida ao integrar a distribuição conjunta a posteriori em relação a σ^2, e vice-versa. Porém, estas integrais não são analíticas, fato que nos conduz a aplicação dos algoritmos MCMC, os quais necessitam das distribuições condicionais completas a posteriori.

3.5. Distribuições Condicionais Completas a Posteriori (DCCP)

Para obter as distribuições em questão, simplesmente assumimos que ao condicionar a posteriori conjunta em relação a um parâmetro, o mesmo passa a ser considerado constante, e como já está explicito no termo de proporcionalidade, este pode ser desconsiderado. Dessa forma, a expressão resultante pode ser denominada de condicional completa a posteriori do outro parâmetro, ou seja, aquele para a qual a condicional completa não foi condicionada.

3.5.1 DCCP para ϕ_i

Neste caso, antes de apresentar a distribuição, vale ressaltar que $(1-\phi_i^2)^{\frac{1}{2}} = \left[(1-\phi_i)(1+\phi_i)\right]^{\frac{1}{2}}$. Assim, é possível obter a partir da expressão (13):

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

$$P(\phi_i|Z_i,\sigma^2) \propto (1-\phi_i)^{\lambda-\frac{1}{2}}(1+\phi_i)^{\alpha-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[h_i\left(\hat{\phi}_i^2 - 2\phi_i\hat{\phi}_i + \phi_i^2\right)\right]\right\} \qquad (14)$$

Nota-se que a expressão (14) não se caracteriza como uma distribuição de probabilidade conhecida, fato este que impõe a aplicação do algoritmo Metropolis-Hastings na geração de valores de ϕ_i que representam amostras da distribuição marginal, a qual não foi obtida diretamente pela inviabilidade de soluções das integrais.

3.5.2. DCCP para σ^2

Um detalhe interessante a ser especificado ao condicionar em relação a ϕ_i, é que para cada série i considerada, tem-se a mesma variância residual σ^2, ou seja, não é considerado o termo σ_i^2. Assim, torna-se necessário à utilização de um produtório (SILVA, 2006) para se condicionar em relação a todas as séries simultaneamente. Da mesma forma que apresentado no item anterior, a partir da expressão (13) obtém-se:

$$P(\sigma^2|Z_i,\phi_i) \propto (\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-(c+1)} \exp\left\{-\frac{h_i(\phi_i-\hat{\phi}_i)^2}{2\sigma^2}-\frac{d}{\sigma^2}\right\}$$

$$P(\sigma^2|Z_i,\phi_i) \propto (\sigma^2)^{-\left(\frac{n}{2}+c+1\right)} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{h_i(\phi_i-\hat{\phi}_i)^2}{2}+d\right)\right\}$$

E considerando agora a utilização do já mencionado produtório, tem-se:

$$P(\sigma^2|Z_i,\boldsymbol{\phi}) \propto \prod_{i=1}^{m}\sigma^{2^{-\left(\frac{n}{2}+c+1\right)}} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{h_i(\phi_i-\hat{\phi}_i)^2+2d}{2}\right)\right\}$$

$$P(\sigma^2|Z_i,\boldsymbol{\phi}) \propto \sigma^{2^{-\left(\frac{n}{2}+c+1\right).m}} \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{m}\left(\frac{h_i(\phi_i-\hat{\phi}_i)^2+2d}{2}\right)\right\}$$

$$P(\sigma^2|Z_i,\boldsymbol{\phi}) \propto \left(\sigma^2\right)^{-\left(\frac{mn}{2}+cm+m-1+1\right)} \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{m}\left(\frac{h_i(\phi_i-\hat{\phi}_i)^2+2d}{2}\right)\right\} \quad (15)$$

De acordo com o aspecto da expressão (15), é possível notar que esta apresenta a f.d.p. de uma distribuição Gama-Inversa, cujos parâmetros são designados por:

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

Dessa forma, temos uma condicional completa representada por uma distribuição de probabilidade conhecida, isto é, $\sigma^2 \mid Z_i,\phi_i \sim GI(c^*,d^*)$. Este fato implica na utilização do algoritmo Gibbs Sampler para gerar amostras da distribuição marginal a posteriori para σ^2.

3.6. Distribuições Preditivas

Sob o enfoque Bayesiano, uma observação futura é descrita por uma distribuição condicional aos dados passados, denominada distribuição preditiva (MIGON; HARRISON, 1985). A distribuição preditiva é obtida pela integral múltipla em relação a todos os parâmetros da distribuição conjunta da observação futura e dos parâmetros, condicionada aos dados passados (de ALBA, 1993).

Segundo Heckman e Leamer (2001), uma forma mais prática consiste em considerar a distribuição preditiva como sendo uma distribuição condicional a posteriori, ou seja, seus valores são atualizados simultaneamente com a geração dos demais parâmetros do modelo via metodologia MCMC. Esta abordagem foi utilizada com sucesso por Silva (2006), portanto a mesma foi considerada no presente trabalho.

A distribuição preditiva de um dado futuro, relacionada com cada indivíduo i listado em uma estrutura de dados em painel, é representada por uma distribuição condicional aos dados passados, ou seja, $P(Z_{i(n+1)} \mid Z_i)$. Esta distribuição é dada por:

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

em que:

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

e $\hat{\phi}_i^* = Z_{i(n+1)} Z_{in} / h_i^*$. Dessa forma, pode se demostrar que: $S_i^{*2} = Z_{i(n+1)}^2 - \frac{Z_{in}^2 Z_{i(n+1)}^2 Z_{in}^2}{Z_{in}^4} = 0.$

Assim, tem-se: $L(Z_{i(n+1)} \propto \exp\left\{-\frac{1}{2\sigma^2}\left[Z_{in}^2\left(\phi_i - \frac{Z_{i(n+1)}Z_{in}}{Z_{in}^2}\right)^2\right]\right\}$. O termo $P(\phi_i,\sigma^2|Z_i,\alpha,\lambda,c,d)$ representa a distribuição conjunta a posteriori expressão (13). Portanto, obtém-se:

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

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

A resolução da integral acima apresenta grande complexidade, fato este que justifica a utilização de métodos MCMC, os quais necessitam da distribuição condicional completa a posteriori para $Z_{i(n+1)}$, a qual é dada por $P(Z_{i(n+1)} \mid Z_i, \sigma^2, \phi_i)$. Assim, para a obtenção desta distribuição condicional, todos os termos de $P(Z_{i(n+1)} \mid Z_i)$ que não contenham $Z_{i(n+1)}$ serão considerados constantes. Dessa forma, é possível obter a seguinte expressão:

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

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

$$P\left(Z_{i(n+1)} \mid Z_i, \sigma^2, \phi_i\right) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[-(Z_{i(n+1)} - \phi_i Z_{in})\right]^2\right\}, \qquad (16)$$

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

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

De forma geral, se constatada a convergência dos algoritmos MCMC, podemos assumir que o conjunto de valores gerados para esta distribuição Normal (16), provenientes de cada q iteração dos algoritmos Metropolis-Hastings e Gibbs Sampler, constituem a distribuição preditiva para um dado futuro, cuja estimativa é $P(\hat{Z}_{(n+1)} \mid Z_i)$. Tal estimativa é representada pela média de todos os valores gerados pela distribuição Normal em questão. Caso seja de interesse, pode-se generalizar esta metodologia para a predição de k dados futuros, porém para esta implementação é necessário obedecer um processo iterativo, fundamentado na ordem de geração dos valores, ou seja, para gerar a distribuição de $Z_{i(n+2)}$, deve-se anteriormente gerar a distribuição de $Z_{i(n+1)}$, e assim sucessivamente até a predição $Z_{i(n+k)}$.

3.7. Implementação dos algoritmos MCMC

Nota-se nos itens 3.4.2 que a distribuição condicional completa para o parâmetro σ^2 é dada por uma distribuição Gama-Inversa, ou seja, ela apresenta uma forma conhecida, portanto passível ao uso do algoritmo Gibbs Sampler.

O mesmo não acontece para a distribuição condicional do parâmetro ϕ_i, o qual não apresenta uma forma definida caracterizada por alguma distribuição de probabilidade conhecida, devendo-se então utilizar, nesta situação, o algoritmo Metropolis-Hastings.

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") para evitar os efeitos dos valores iniciais adotados. A constatação final da convergência foi realizada por meio dos critérios de Geweke (1992) e de Raftery e Lewis (1992). Nogueira (2004) recomenda a utilização de diferentes critérios para onstatar a convergência. Ambos os critérios são avaliados mediante o pacote BOA ("Bayesian Output Analysis") do software R.

Os códigos utilizados para a implementação do programa no software R são apresentados no Apêndice 2.

3.8. Comparação de modelos: Fator de Bayes e Capacidade Preditiva

No presente estudo, foram considerados três diferentes modelos, dados por:

Modelo 1: priori simétrica, $\mathrm{P}(\phi_i \mid \alpha, \lambda) = \mathrm{Beta\ Gen}(13,5;5)$

Modelo 2: priori assimétrica, $\mathrm{P}(\phi_i \mid \alpha, \lambda) = \mathrm{Beta\ Gen}(4;2)$;

Modelo 3: priori constante, $\mathrm{P}(\phi_i \mid \alpha, \lambda) = \mathrm{Beta\ Gen}(4;2)$

A comparação entre estes modelos foi realizada via Fator de Bayes (FB) sob o enfoque apresentado por Barreto e Andrade (2004). Este fator utiliza valores gerados pelos métodos MCMC para obter as estimativas do fator de normalização, $P(Z \mid M_p)$, também denominado de Verossimilhança Marginal, o qual compõe a expressão do Fator de Bayes:

$$\mathrm{FB}_{ij} = \frac{\hat{\mathrm{P}}(\boldsymbol{Z} \mid \mathrm{M}_i)}{\hat{\mathrm{P}}(\boldsymbol{Z} \mid \mathrm{M}_j)} = \frac{\frac{1}{\mathrm{Q}}\sum_{q=1}^{\mathrm{Q}} \mathrm{L}(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(q)}, \mathrm{M}_i)}{\frac{1}{\mathrm{Q}}\sum_{q=1}^{\mathrm{Q}} \mathrm{L}(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(q)}, \mathrm{M}_j)}.$$

Como frisado anteriormente, $\boldsymbol{\theta}^{(q)}$ indica os valores gerados para os parâmetros na q-ésima iteração (q = 1,2,...,Q) para cada um dos modelos comparados. Assim, o termo $L(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(\theta)}, M_p)$ corresponde a valores da função de verossimilhança obtidos pela substituição dos valores atuais dos parâmetros gerados pelos algoritmos MCMC.

Usando a função de verossimilhança adotada neste estudo, tem-se a seguinte estimativa da Verossimilhança Marginal de um modelo p:

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

$$\hat{P}(\mathbf{Z}|M_p) = \frac{1}{Q}\sum_{q=1}^{Q}\left[\prod_{i=1}^{m}\left(\frac{1}{\sigma^{2(q)}}\right)^{\frac{n}{2}} .\left(1-\phi_i^{2(q)}\right)^{\frac{1}{2}} .\exp\left\{-\frac{1}{2\sigma^{2(q)}}\left[\mathrm{S}_{\mathrm{i}}^2 + \mathrm{h}_{\mathrm{i}}\left(\hat{\phi}_i^2 - 2\phi_i^{(q)}\,\hat{\phi}_i + \phi_i^{2(q)}\right)\right]\right\}\right]$$

As comparações dos modelos via Fator de Bayes foram estruturadas de acordo com a Tabela 1.

Tabela 1 – Esquema de comparação das prioris consideradas

Modelos (Prioris)	**Critério**
Modelo 1 x Modelo 2	$\mathrm{FB}_{12} = \dfrac{\hat{\mathrm{P}}(\boldsymbol{Z}\mid \mathrm{M}_1)}{\hat{\mathrm{P}}(\boldsymbol{Z}\mid \mathrm{M}_2)}$
Modelo 1 x Modelo 3	$\mathrm{FB}_{13} = \dfrac{\hat{\mathrm{P}}(\boldsymbol{Z}\mid \mathrm{M}_1)}{\hat{\mathrm{P}}(\boldsymbol{Z}\mid \mathrm{M}_3)}$
Modelo 2 x Modelo 3	$\mathrm{FB}_{23} = \dfrac{\hat{\mathrm{P}}(\boldsymbol{Z}\mid \mathrm{M}_2)}{\hat{\mathrm{P}}(\boldsymbol{Z}\mid \mathrm{M}_3)}$

Para avaliar a capacidade preditiva de cada modelo, ou seja, verificar os valores preditos para um dado futuro por meio de suas distribuições preditivas, 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 serão estimados sem a presença destas observações, as quais serão preditas pela metodologia adotada. Uma forma prática e precisa de avaliar a eficiência destas predições, está relacionada com o fato dos intervalos de credibilidade conterem os verdadeiros valores ocultados na análise. Portanto, de acordo com Silva (2006), a porcentagem de séries cujos intervalos de credibilidade contém o verdadeiro valor da última observação pode ser considerada uma boa medida para a capacidade preditiva do modelo.

3.9. Dados simulados

O processo de 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 desconsiderada para avaliação da Capacidade Preditiva. A Figura 4 apresenta esquematização deste processo.

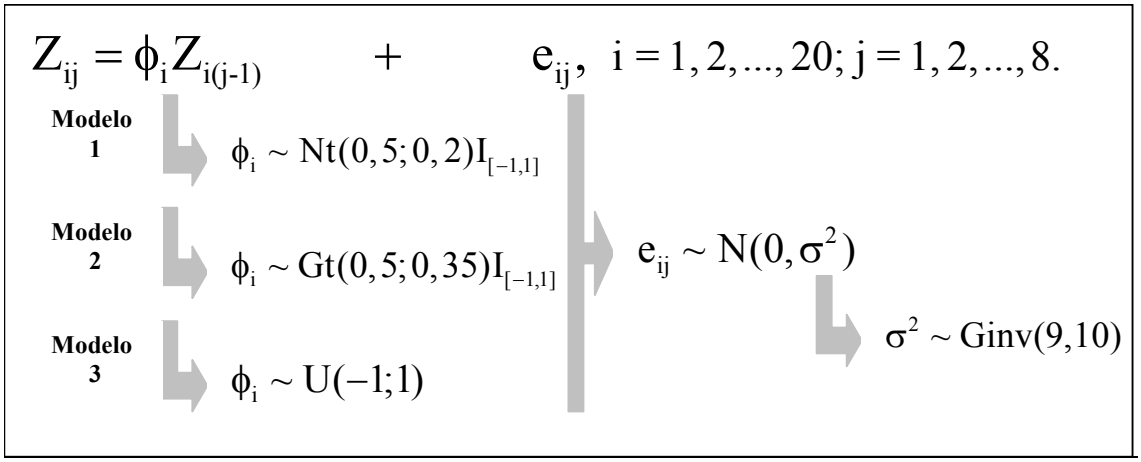


Figura 4 – Diagrama que descreve o processo de simulação (Nt: Normal truncada, Gt: Gumbel truncada, U: Uniforme, N: Normal, Ginv: Gama Inversa).

O procedimento apresentado na Figura 4 foi executado uma única vez, ou seja, foi realizada apenas uma única simulação para cada modelo estudado, e isto se deve ao fato da Inferência Bayesiana não necessitar de repetições de experimentos para avaliar a adequacidade de modelos, uma vez que conceitos freqüentistas como Erro Quadrático Médio e Viés geralmente não se aplicam a abordagem Bayesiana.

Na análise dos dados simulados, não se objetivou comparar os modelos 1, 2 e 3 (item 3.6), apenas objetivou-se a obtenção de um resultado geral de validação de cada modelo separadamente. Para tanto avaliou-se a qualidade de estimação dos parâmetros ϕ_i , σ^2 e da última observação , ou seja, se os verdadeiros valores (valores paramétricos da simulação) encontravam-se realmente contidos no intervalo de credibilidade de 95%. A comparação formal por Fator de Bayes sob esta condição de simulação não foi

considerada relevante, pois, por exemplo, ao simular valores de ϕ_i por uma distribuição simétrica, não foram encontradas razões para a qual uma distribuição assimétrica poderia ser indicada como a melhor.

3.10. Dados Reais

Sucintamente, a técnica de microarray consiste na deposição de seqüências de cDNA (fragmentos de DNA complementar, que são clonados em bactérias e estão na forma de fita, ao invés de estar na forma de dupla hélice) 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 (ESTEVES, 2007).

Suponha que se queira testar quais genes atuam sobre um processo celular qualquer, como resistência à doença e a condições ambientais extremas, dentre outras. Deste modo, extrai-se o RNA total do organismo na presença e na ausência da doença, ou 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 expressos no organismo.

Os dados originais de um experimento de microarray são imagens que representam os níveis de expressão dos genes fixados no substrato, que são analisadas por um software específico que gera uma tabela de dados numéricos contendo os valores de intensidade de emissão de luz (geralmente em nm) de cada gene em cada situação experimental considerada (FACELI, 2005). Em relação aos dados MTS, estes são caracterizados pela repetição destes procedimentos ao longo do tempo, de forma que se torna possível avaliar o nível de expressão em diferentes instantes.

Os dados de expressão geralmente apresentam uma distribuição normal depois de uma transformação logarítmica, e por isso esse tipo de transformação é muito comum nos dados de microarray (HARIHARAN, 2003). Além da distribuição se tornar aproximadamente normal após esse tipo de transformação, ela auxilia na visualização

dos dados e faz com que os níveis de expressão dos genes hiper e hipo-expressos, resultantes de experimentos competitivos, tenham mesma magnitude e sinais opostos (QUACKENBUSH, 2001).

Os dados utilizados no presente estudo são referentes à expressão gênica do ciclo celular de células HeLa (células humanas epiteliais provenientes da fase final de crescimento), e foram coletados por Whitfield et al. (2002) e analisados por Fujita et al. (2007). Esses dados contem médias de três ciclos completos de 16 horas das células HeLa, com expressões avaliadas em intervalos de uma hora. 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), e, além disso, foram utilizadas apenas as últimas 8 observações de cada gene. Esta adaptação foi adotada para caracterizar a situação típica de dados em painel, na qual se tem maior número de séries que número de observações por série, pois a última observação temporal de cada gene foi excluída da análise para verificar a capacidade preditiva, fato que resultou em um arquivo final com 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/. Neste, a variável dependente é dada pela intensidade de expressão transformadas para a escala logarítmica.

4. RESULTADOS E DISCUSSÃO

4.1. Dados simulados

As Tabelas 2, 3 e 4 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 constante.

É possível notar em todas estas Tabelas que realmente a metodologia Bayesiana apresentada foi eficiente, pois as estimativas obtidas assumiram valores bem próximos dos verdadeiros, como era de se esperar, uma vez que as distribuições a priori usadas representaram quase que perfeitamente aquelas utilizadas para a geração dos valores paramétricos.

Nota-se, também, que para todos os parâmetros, os verdadeiros valores encontram-se dentro do intervalo de credibilidade de 95%. Este é um ponto relevante para discussões relacionadas com a teoria de propriedades de estimadores Bayesianos, pois quando o intervalo não contém o valor paramétrico, estando este abaixo ou acima do limite inferior, pode-se dizer, respectivamente, que o mesmo apresenta a característica de subestimação ou de superestimação (SILVA, 2006).

Dessa forma é possível constatar que realmente a utilização de distribuições a priori Beta para representar outras distribuições é eficiente, e representa uma forma mais fácil de se testar diferentes distribuições a priori, visto que o processo algébrico de obtenção das distribuições condicionais completas a posteriori, necessárias para implementação dos algoritmos MCMC foi o mesmo, e alteraram-se apenas os valores dos parâmetros da distribuição Beta em cada caso.

Tabela 2 – Valores paramétricos (ϕ_i) simulados da Normal, estimativa dada pela média da 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 & Lewis (FD_{RL}) ao utilizar a priori Beta simétrica

Série	ϕ_i	$\hat{\phi}_i$	**Li**	**Ls**	$\mathbf{P_G}$	$\mathbf{FD_{RL}}$
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 3 – Valores paramétricos (ϕ_i) simulados da Gumbel, estimativa dada pela média da 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 & Lewis (FD_{RL}) ao utilizar a priori Beta assimétrica

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 4 – Valores paramétricos (ϕ_i) simulados da Uniforme, estimativa dada pela média da 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 & Lewis (FD_{RL}) ao utilizar a priori Beta constante

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

Em recentes estudos envolvendo análise Bayesiana de modelos de séries temporais (SILVA, 2006) e de modelos de regressão não-linear (SILVA, 2006; SAVIAN, 2008) a técnica de simulação de dados foi utilizada para verificar a eficiência da metodologia e também, segundo estes autores, para validar o recurso computacional utilizado. Este último fato é importante, porque os programas elaborados em softwares estatísticos como o R e o SAS estão sujeitos a presença de erros, comumente chamados de BUG, uma vez que as distribuições condicionais completas geralmente representam funções complexas, com muitos índices e grande demanda de operações matriciais. Assim, diante dos resultados das Tabelas 2, 3 e 4, mesmo considerando uma situação ótima, na qual se utilizaram distribuições a priori apropriadas, caso o programa apresentasse algum problema de ordem técnica, os resultados certamente não teriam sido tão fiéis em relação aos valores paramétricos.

Quanto à verificação da convergência dos algoritmos MCMC, nota-se que o critério de Geweke (1992) apenas rejeitou a hipótese de convergência ($P_G < 0,05$) em três situações, série 7 da Tabela 2 , série 8 da Tabela 3 e séries 12 e 13 da Tabela 4, enquanto que o critério de Raftery e Lewis (1992), fundamentado no fator de dependência que indica convergência para valores menores que 5, mostrou que todas as cadeias convergiram. Assim, concluiu-se que não era necessário executar novamente os programas com um número maior de iterações, pois nos piores cenários apontados pelo critério de Geweke, o critério de Raftery e Lewis indicou convergência. Portanto, se ambos os critérios tivessem apontado falta de convergência para um mesmo parâmetro, então seria necessário efetuar novamente o processo de amostragem MCMC utilizando um número maior de iterações.

As Tabelas 5, 6 e 7 apresentam as estimativas para a última observação de cada série por meio da distribuição preditiva, respectivamente para cada distribuição utilizada como priori, ou seja, simétrica, assimétrica e constante.

De forma geral, todas as distribuições a priori utilizadas providenciaram bons resultados relacionados com a predição de um único valor futuro, principalmente a distribuição a priori simétrica (Tabela 5), a qual apresentou uma capacidade 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. As distribuições a priori assimétricas (Tabela 6) e constante (Tabela 7) apresentaram, respectivamente, capacidades preditivas de 80 e 90%, pois na Tabela 6 para quatro séries (4,5,6 e 7) e na Tabela 7 para duas séries (8 e 18) os valores paramétricos não estavam inclusos no intervalo.

Tabela 5 – Última observação (Z_{i8}), estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar a priori Beta simétrica

Série	$\mathbf{Z_{i8}}$	$\mathbf{\hat{Z}_{i8}}$	**Li**	**Ls**	$\mathbf{P_G}$	$\mathbf{FD_{RL}}$
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 6 – Última observação (Z_{i8}), estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar a priori Beta assimétrica

Série	$\mathbf{Z_{i8}}$	$\mathbf{\hat{Z}_{i8}}$	**Li**	**Ls**	$\mathbf{P_G}$	$\mathbf{FD_{RL}}$
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 7 – Última observação (Z_{i8}), estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar a priori Beta constante

Série	$\mathbf{Z_{i8}}$	$\mathbf{\hat{Z}_{i8}}$	**Li**	**Ls**	$\mathbf{P_G}$	$\mathbf{FD_{RL}}$
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

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 Normal, e o mesmo é válido para a assimétrica e a constante, nas situações em que os parâmetros foram gerados respectivamente por uma Gumbel e uma Uniforme. Resumidamente, os resultados das Tabelas 5, 6 e 7 indicam que a distribuição preditiva apresenta-se como uma ferramenta eficiente para a predição de valores futuros, visto que ao se considerar conjuntamente todas as Tabelas, temos uma capacidade preditiva de 90%. Silva (2006) obteve resultados semelhantes ao calcular a capacidade preditiva, a qual foi de 80%, considerando as distribuições a priori t-Student multivariada, Normal-Multivariada e Priori de Jeffreys.

A Tabela 8 apresenta as estimativas da variância residual para cada distribuição a priori considerada.

Tabela 8 - Variância residual paramétrica (σ^2), estimativa dada pela média da 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 & Lewis (FD_{RL})

Priori	σ^2	$\hat{\sigma}^2$	**Li**	**Ls**	P_G	FD_{RL}
Simétrica	1,8207	2,2567	1,2742	3,9379	0,8478	0,9938
Assimétrica	2,3978	3,1094	1,7475	5,4065	0,3775	1,0018
Constante	1,1494	1,9220	1,0886	3,3370	0,9674	0,9698

Quanto às estimativas apresentadas na Tabela 8, 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 2 a 6, todos os intervalos de credibilidade contêm 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.

Para ilustrar a análise das cadeias geradas pelo algoritmo Gibbs Sampler, os gráficos com os valores gerados a cada iteração e a densidade da distribuição a posteriori para o parâmetro variância do erro são apresentados na Figura 2.

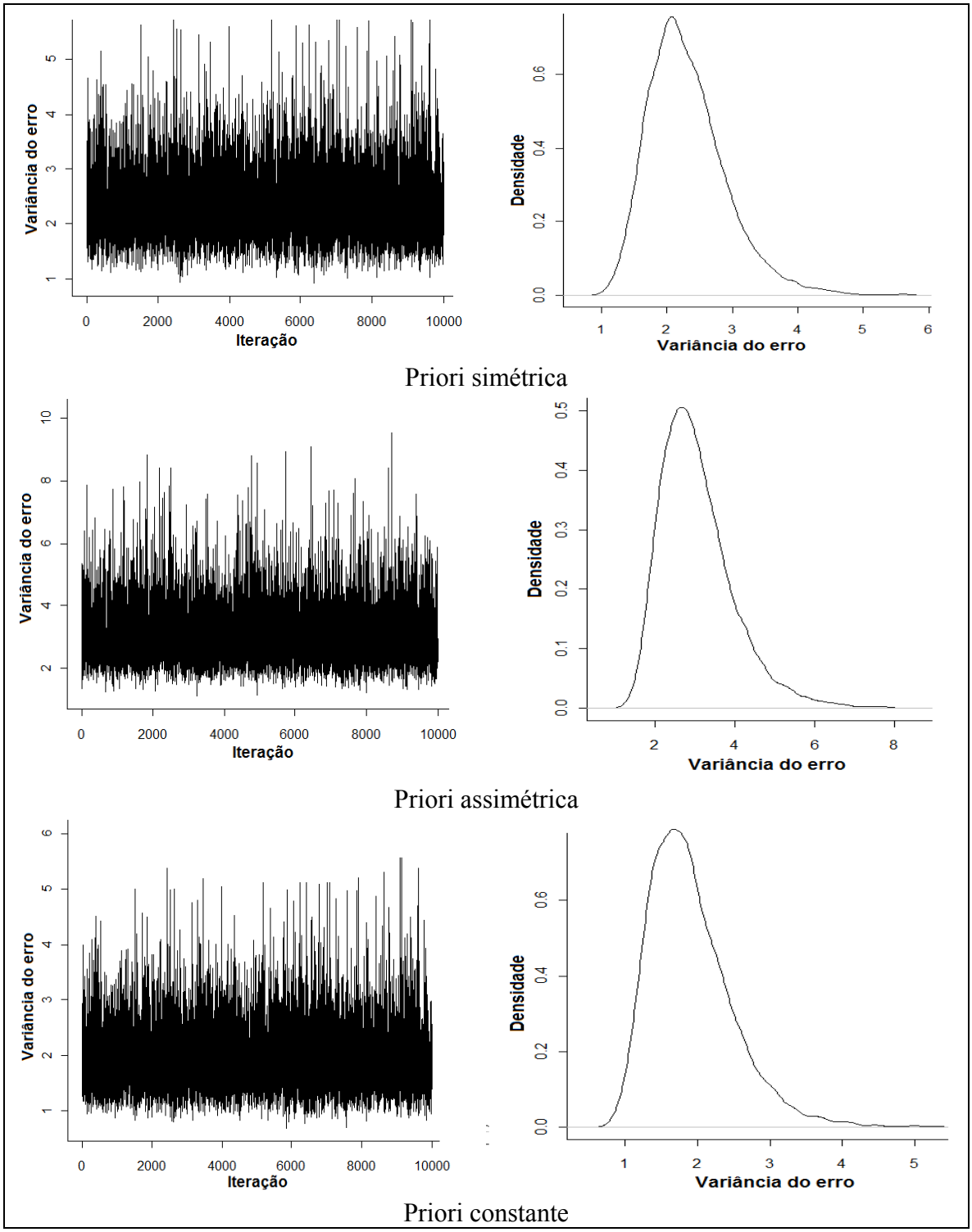


Figura 2 – Valores gerados para a variância do erro a cada iteração e respectiva densidade da distribuição marginal a posteriori para cada priori utilizada.

4.2. Dados reais

As Tabelas 9,10 e 11 apresentam as estimativas dos parâmetros do modelo AR (1) para dados em painel, respectivamente para cada distribuição utilizada como distribuição a priori (simétrica, assimétrica e constante) ajustadas ao mesmo conjunto de dados de expressão gênica avaliada ao longo do tempo.

De acordo com os resultados apresentados nas Tabelas 9, 10 e 11, é possível inferir que as séries referentes as estimativas obtidas para o coeficiente de auto-regressão (ϕ_i) foram similares em relação as distribuições a priori utilizadas. Nota-se também, que as distribuições a priori simétrica e constante (Tabelas 9 e 11) providenciaram uma porcentagem de significância um pouco menor que aquela obtida para a distribuição a priori assimétrica (Tabela 10), uma vez que para esta última o intervalo de credibilidade não conteve zero pra 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).

Tabela 9 – Estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar distribuição a priori Beta 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

Tabela 10 – Estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar distribuição a priori Beta 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

Tabela 11 – Estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar distribuição a priori Beta constante

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

Outro ponto relevante quanto às estimativas mostradas nas Tabelas 9, 10 e 11 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 são 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.

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)

As Tabelas 12, 13 e 14 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 constante ao se considerar os dados reais de expressão gênica ao longo do tempo.

Tabela 12 – Estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar distribuição a priori Beta 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

Tabela 13 – Estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar distribuição a priori Beta 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

Tabela 14 – Estimativa dada pela 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 & Lewis (FD_{RL}) ao utilizar a priori Beta constante

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 nas Tabelas 12, 13 e 14 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. Vários autores que também avaliaram a capacidade preditiva de modelos auto-regressivos sob enfoque Bayesiano mediante intervalos de credibilidade a posteriori, também apontaram para este sucesso no procedimento de previsão. Dentre estes, pode-se citar de Alba (1993), que ao simular quatro séries independentes sob um modelo auto-regressivo de ordem quatro, AR (4), obteve 75% de eficiência na predição de um dado futuro, e Silva (2006), que ao ajustar o modelo AR (2) para dados em painel a observações temporais de valores genéticos de touros Nelore obteve 85% de acerto na previsão de um dado futuro com a utilização da priori t-multivariada e 77,78% com a utilização da distribuição a priori normal multivariada.

Na prática, a aplicação desta metodologia de previsão a 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 a 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 contendo as amplitudes dos intervalos de credibilidade da distribuição a posteriori para cada priori adotada, o qual é mostrado na Figura 3.

A Tabela 15 apresenta as estimativas da variância residual para cada distribuição a priori considerada.

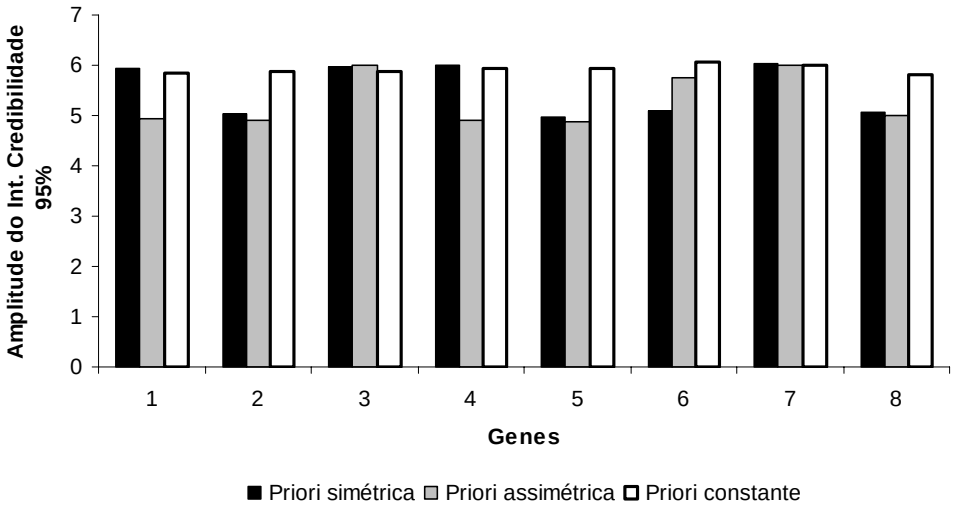


Figura 3 – Amplitude do intervalo de credibilidade de 95% para a última observação de cada série correspondente a expressão gênica de células HeLa.

Tabela 15 – 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 & Lewis (FD_{RL})

Priori	$\hat{\sigma}^2$	**Li**	**Ls**	P_G	FD_{RL}
Simétrica	2,5911	0,9303	5,3317	0,1339	1,0101
Assimétrica	2,1582	0,9219	5,3534	0,0741	1,0266
Constante	2,6467	0,9445	5,3087	0,4439	1,0184

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 15 que a estimativa da variância residual para esta mesma 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 a priori como sendo a melhor, segundo Silva (2006), 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 de modelos, como por exemplo, o Fator de Bayes.

Para ilustrar a análise das cadeias geradas pelo algoritmo Gibbs Sampler na análise dos dados reais, os gráficos com os valores gerados a cada iteração e a densidade da distribuição a posteriori para o parâmetro variância do erro são apresentados na Figura 3.

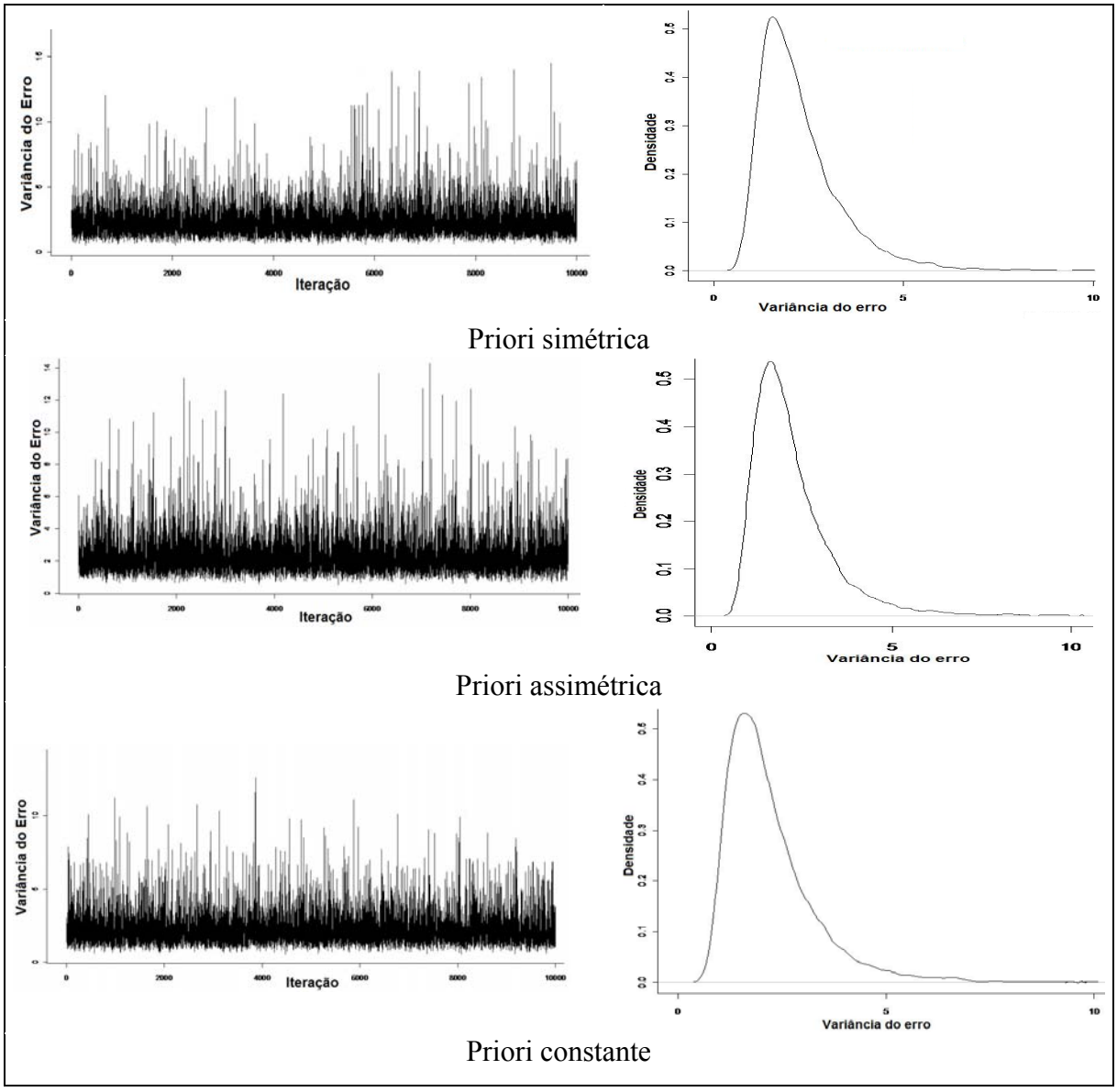


Figura 3 – Valores gerados para a variância do erro a cada iteração e respectiva densidade da distribuição marginal a posteriori para cada priori utilizada.

Na Tabela 16 são mostrados os valores obtidos para o Fator de Bayes correspondentes a comparação das distribuições a priori ao se analisar os dados temporais de expressão gênica.

Tabela 16 – Comparação das distribuições a priori por meio do Fator de Bayes considerando os dados reais

Modelos (Prioris)	**Critério**
Modelo 1 (P. simétrica) x Modelo 2 (P. assimétrica)	$FB_{12} = \frac{1,4667x10^{-7}}{1,5115x10^{-7}} = 0,9704$
Modelo 1 (P. simétrica) x Modelo 3 (P. constante)	$FB_{13} = \frac{1,4667x10^{-7}}{1,4981x10^{-7}} = 0,9790$
Modelo 2 (P. assimétrica) x Modelo 3 (P. constante)	$FB_{23} = \frac{1,5115x10^{-7}}{1,4981x10^{-7}} = 1,0089$

Os resultados contidos na Tabela 16 informam que a distribuição a priori assimétrica proporcionou uma melhor qualidade de ajuste em relação às outras duas, embora as diferenças entre elas sejam praticamente irrisórias. Estas afirmações dizem respeito ao fato do fator de Bayes ser menor que a unidade quando se comparou a distribuição a priori simétrica com a assimétrica, estando esta última no denominador, e maior que a unidade, quando se comparou a assimétrica com a constante, estando esta no numerador. Quanto à comparação entre simétrica e constante, esta última apresentou maior qualidade de ajuste, uma vez que foi considerada no denominador e o Fator de Bayes foi menor que a unidade.

De forma geral, como já comentado no parágrafo anterior, as diferenças entre estes valores são muito pequenas, o que na prática impossibilita afirmar que a distribuição a priori assimétrica deve ser indicada como a melhor para estudos que envolvam ajuste do modelo AR (1) para dados em painel a observações temporais, de expressão gênica por microarray. Porém, um fator relevante nesta análise é a apresentação deste esquema de comparação via Fator de Bayes, o qual envolveu a obtenção de verossimilhanças marginais via cadeias geradas por MCMC, evitando assim o envolvimento com métodos complexos de resoluções de integrais múltiplas que muitas vezes inviabilizam a utilização deste método de comparação de modelos.

Os gráficos contendo os valores observados e estimados para cada série de expressão gênica, ao considerar a distribuição a priori assimétrica, 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.

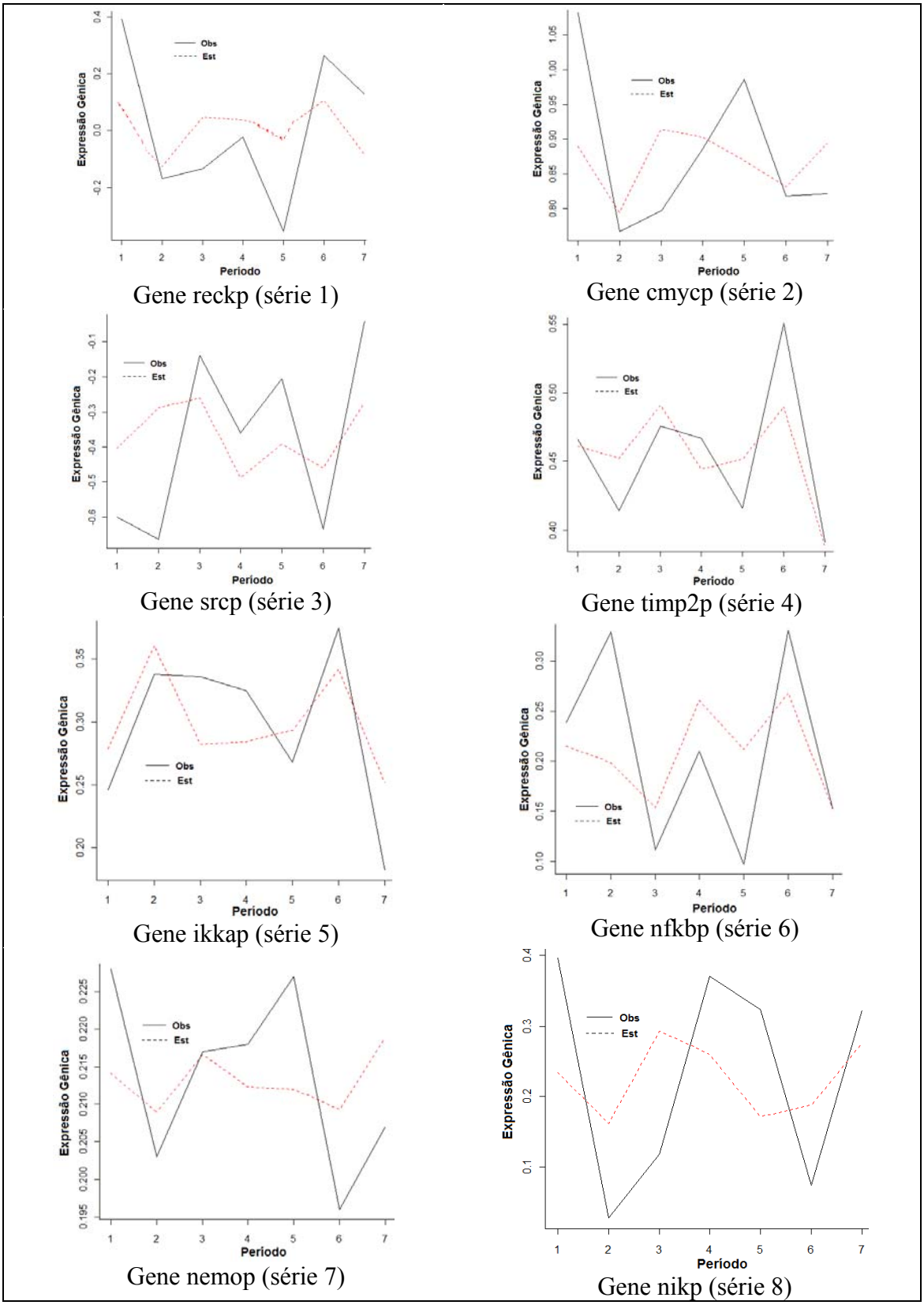


Figura 4 – Séries temporais de expressão gênica observadas (Obs) e estimadas (Est) pelo modelo auto-regressivo de primeira ordem, AR (1), para dados em painel considerando a priori assimétrica.

5. CONCLUSÕES

1) Todos os passos para a realização da análise Bayesiana do modelo auto-regressivo de primeira ordem, AR (1), para dados em painel foram executados com sucesso, pois por meio do teorema de Bayes, foi possível considerar a distribuição de verossimilhança exata e diferentes distribuições a priori dadas por diferentes distribuições Beta. Além disso, a metodologia empregada possibilitou utilizar, de forma prática e objetiva, o método de comparação de modelos pelo Fator de Bayes via cadeias geradas pelos algoritmos MCMC.
2) 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.
3) 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.
4) A comparação de modelos indicou numericamente, mediante valores do Fator de Bayes, 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.

6. REFERÊNCIAS

ARELLANO, M.; BOVER, O. La econometría de datos de panel. **Investigaciones Econômicas,** Madrid, Spain, v. 14, n. 1, p. 3-45, 1990.

BALTAGI, B. **Econometric analysis of panel data**. 2. ed. New York, USA: John Wiley and Sons, 2000. 314 p.

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, Mar. 2004.

BEARZOTI, E. **Introdução à estatística matemática**. Lavras, MG: Editora FAEPE-UFLA, 1998. (Apostila).

de ALBA, E. Constrained forecasting in autoregressive time series models: A Bayesian analysis. **International Journal of Forecasting**, New York, v. 9, n. 1, p. 95-108, Apr. 1993.

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** – Relatório técnico 250. São Carlos, SP: ICMC, 2005.

FALK, B.; ROY, A. Forecasting using the trend model with autoregressive errors. **International Journal of Forecasting**, London, v. 21, n. 4, p. 291-302, Oct./Dec. 2005.

FREES, E. **Longitudinal and panel data**. Cambridge, UK: CambridgeUniversity Press, 2004. 320 p.

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.

FUJITA, A. **Análise de dados de expressão gênica**: normalização de microarrays e modelagem de redes regulatórias. 2007. 90 f. Tese (Doutorado em Bioinformática) – Universidade de São Paulo, São Paulo, 2007.

GELFAND, A.E. Gibbs sampling. **Journal of the American Statistical Association**, v. 95, p. 1300-1304, 2000.

GELMAN, A.; RUBIN, D.B. Inference from iterative simulation using multiple sequence. **Statistical Science**, Hayward, v. 7, n. 4, p. 457-511, May 1992.

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. p. 625-631.

GHOSH, M.; HEO, J. Default Bayesian priors for regression models with first-order autoregressive residuals. **Journal of Time Series Analysis**, New York,v. 24, n. 3, p. 269-282, May 2003.

HARIHARAN, R. The analysis of microarray data. **Pharmacogenomics,** v. 4, n. 4, p. 477-497, 2003.

HASTINGS, W.K. Monte Carlo sampling methods using markov chains and their applications. **Biometrika**, v. 57, p. 97-109, 1970.

HECKMAN, J.; LEARMER, E.**Handbook of econometrics.** Amsterdam, Netherlands: Elsevier Science, 2001. v. 5, p. 744.

HEIDELBERGER, P.; WELCH, P. Simulation run length control in the presence of an initial transient. **Operations Research**, Maryland, v. 31, n. 6, p. 1109-44, Nov.-Dec. 1983.

HIRANO, K. Semiparametric bayesian inference in autoregressive panel data models. **Econometrica**, Tel Aviv, v. 70, n. 2, p. 781-799, Mar. 2002.

HSIAO, C.; SUN, B.H. To pool or not to pool panel data. Panel data econometrics: future directions. In: KRISHNAKUMAR, J.; RONCHETTI, E. (Eds.). **Papers in honour of professor Pietro Balestra**. Amsterdam: North Holland, 2000. p. 881-899, 2000.

JEFFREYS, H. **Theory of probability**. Oxford, UK: Claredon Press, 1961.

JONG, H. Modeling and simulation of genetic regulatory systems: a literature review. **Journal of Computational Biology**, v. 9, n. 1, p. 69-105, 2002.

KASS, R.E.; RAFTERY, A.E. Bayes factors. **Journal of the American Statistical Association**, Alexanria, v. 90, n. 430, p. 773-795, Jun. 1995.

KITTEL, B. Sense and sensitivity in pooled analysis of political data. **European Journal of Political Research**., v. 35, p. 225-253, 1999.

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. 60 f. Dissertação (Mestrado) – Universidade de São Paulo, 2007.

M METROPOLIS, N.; A.W.; ROSENBLUTH, M.N.; ROSENBLUTH, A.H.; TELLER, TELLER, E. Equations of State Calculations by Fast Computing Machines. **Journal of Chemical Physics**, v. 21, p. 1087-1092, 1953.

MIGON, H. S.; HARRISON, P. J. An application of non-linear Bayesian forecasting to television advertising. In: BERNARDO, J. M. et al. (Eds.). **Bayesian statistics.** 2. ed. New York: Oxford University Press, 1985. p. 681-696.

MORAIS,T.S.S; SILVA, F.S.; SILVEIRA, F.G.**Utilização da Priori Beta(P,Q) para os Parâmetros do Modelo Randon Coefficient First Order Autoregressive – RCA(1)**. 53ª RBRAS- Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria de 2008. (Participações em eventos/Congresso).

MORETTIN, P.A.; TOLOI, C.M.C. **Análise de séries temporais**. São Paulo: Edgard Blucher, 2004. 535 p.

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

NI, S.; SUN, D. Noninformative priors and frequentist risks of Bayesian estimators of vector-autoregressive models. **Journal of Econometrics**, New York, v. 115, n. 1, p. 159-197, jul. 2003.

NOGUEIRA, D. A. **Proposta e avaliação de critérios de convergência para o método de Monte Carlo via Cadeias de Markov**: casos uni e multivariados. 2004. 142 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2004.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Editora Fundação Calouste Gulbenkian, 2003.

QUACKENBUSH, J. Computational analysis of microarray data. **Nature Reviews Genetics,** v. 2, n. 6, p. 418-427, 2001.

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. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. **Biometrika**, v. 83, p. 251-266, 1996.

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.

REIS, R. L.; MUNIZ, J.A.; SILVA, F. F.; SÁFADI, T.; AQUINO, L.H. Inferência bayesiana na análise genética de populações diplóides: estimação do coeficiente de endogamia e da taxa de fecundação cruzada. **Ciência Rural**, v. 38, 2008. No prelo.

ROSA, G.J.M. **Análise bayesiana de modelos mistos robustos via amostrador de Gibbs.** 1998. 57 f. Tese (Doutorado em Estatística e Experimentação Agronômica) – Universidade de São Paulo, Piracicaba, 1998.

SAFADI, T.; MORETTIN, P.A. Bayesian analysis of autoregressive models with normal randon coefficients. **Journal of Statistical Computation and Simulation**, Blaksburg, v. 73, n. 8, p. 563-573, Aug. 2003.

SAVIAN, T.V. Análise bayesiana para modelos de degradabilidade ruminal. 2008. 81 f. Tese (Doutorado) – Universidade Federal de Lavras, Lavras, 2008.

SCOLFORO**,** J.R.S. **Mensuração florestal**: módulo 6. Modelos de crescimento e produção – Parte 2. Lavras, MG: ESAL/FAEPE, 1995**.** 243 p.

SILVA, F.F. **Análise bayesiana do modelo auto-regressivo para dados em painel**: aplicação na avaliação genética de touros da raça Nelore. 2006. 100 f. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2006.

SILVA, N.A.M.; MUNIZ, J.A.; SILVA, F. F.; AQUINO, L.H. Estudo de parâmetros de crescimento de bezerros nelore por meio de um modelo de regressão linear: uma abordagem bayesiana. **Ciência Animal Brasileira** (UFG), v. 7, p. 57-65, 2006.

SILVA, C.H.O. **A proposed framework for establishing optimal genetic designs for estimating narrow-sense heritability**. 2000. 98 f. 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.

TURKMAN, M.A.A. **Introdução aos métodos bayesianos**. Disponível em: <http://www.deio.fc.ul.pt/getfile.asp?id=766>. Acesso em: set. 2008.

VERMAAK, J.; ANDRIEU, C.; DOUCET, A.; GOSILL, S.J. **Bayesian model selection of autoregressive processes**. Cambridge: Cambridge University Engineering Department, 2000. (Technical Report CUED/F-INFENG/TR, 360).

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.

YAMAGUCHI R, YOSHIDA R, IMOTO S, HIGUCHI T, MIYANO S. Finding module-based gene networks in time-course gene expression data with state space models. **IEEE Signal processing magazine**. 24:37-46 2007.

YANG, R.Y.; BERGER, J.O. Estimation of a covariance matrix using the reference prior. **Annals of Statistics**, Minneapolis, v. 22, n. 3, p. 1195-1211, Sept. 1994.

ZELLNER, A. Models, prior information, and bayesian analysis. **Journal of Econometrics**, New York, v. 75, n. 1, p. 51-68, nov. 1996.

ZHOU, Y.Y.; ROY, A. Effect of tapering on accuracy of forecasts made with stable estimators of vector autoregressive processes. **International Journal of Forescasting**, London, v. 22, n. 1, p. 169-180, jan./mar. 2006.

APÊNDICE

APÊNDICE 1 – Função de Verossimilhança

$$P(Z_i^* \mid \phi_i, \sigma^2) = \prod_{j=2}^{n} P(Z_{ij}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} .\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=2}^{n}\left(Z_{ij} - \phi_i Z_{i(j-1)}\right)^2\right\}$$

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

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

$$P(Z_{i1} \mid \phi_i, \sigma^2) \propto \left(\frac{1-\phi_i^2}{\sigma^2}\right)^{\frac{1}{2}} .\exp\left\{-\frac{1}{2\sigma^2}\left(Z_{i1}^2\left(1-\phi_i^2\right)\right)\right\} \text{ Moretin e Toloi}$$

$$P(Z_i \mid \phi_i, \sigma^2) = P(Z_i^* \mid \phi_i, \sigma^2).P(Z_{i1} \mid \phi_i, \sigma^2)$$

$$P(Z_i \mid \phi_i, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{\frac{n-1}{2}+\frac{1}{2}} .\left(1-\phi_i^2\right)^{\frac{1}{2}} .\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=2}^{n} Z_{ij}^2 - 2\phi_i\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)} + \phi_i^2\sum_{j=2}^{n} z_{i(j-1)}^2 + Z_{ij}^2 - Z_{i1}^2\phi_i^2\right\}$$

$$P(Z_i \mid \phi_i, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} .\left(1-\phi_i^2\right)^{\frac{1}{2}} .\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{n} Z_{ij}^2 - 2\phi_i\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)} + \phi_i^2\left(\sum_{j=2}^{n} Z_{i(j-1)}^2 - Z_{i1}^2\right)\right)\right\}$$

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

$$P(Z_i \mid \phi_i, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} .\left(1-\phi_i^2\right)^{\frac{1}{2}} .\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{j=1}^{n} Z_{ij}^2 - \frac{\sum_{j=2}^{n-1} Z_{ij}^2\left(\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)}\right)}{\left(\sum_{j=2}^{n-1} Z_{ij}^2\right)^2} + \frac{\sum_{j=2}^{n-1} Z_{ij}^2\left(\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)}\right)}{\left(\sum_{j=2}^{n-1} Z_{ij}^2\right)^2} - 2\phi_i\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)} + \phi_i^2\sum_{j=2}^{n-1} Z_{ij}^2\right]\right\}$$

$$P(Z_i \mid \phi_i, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} .\left(1-\phi_i^2\right)^{\frac{1}{2}} .\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{j=1}^{n} Z_{ij}^2 - \sum_{j=1}^{n-1} Z_{ij}^2\frac{\left(\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)}\right)^2}{\left(\sum_{j=2}^{n-1} Z_{ij}^2\right)^2} + \sum_{j=1}^{n-1} Z_{ij}^2\left(\frac{\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)}}{\left(\sum_{j=2}^{n-1} Z_{ij}^2\right)^2} - \frac{2\phi_i\sum_{j=2}^{n} Z_{ij}Z_{i(j-1)}}{\sum_{j=2}^{n-1} Z_{ij}^2} + \phi_i^2\right)\right]\right\}$$

Denominando:

$$h_i = \sum_{j=2}^{n-1} Z_{ij}^2 \quad , \hat{\phi}_i = \frac{\sum_{j=2}^{n} Z_{ij} Z_{i(j-1)}}{\sum_{j=2}^{n-1} Z_{ij}^2} \quad , S_i^2 = \sum_{j=1}^{n} Z_{ij}^2 - h_i \hat{\phi}_i^2$$

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

$$P(Z_i \mid \phi_i, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \left(1-\phi_i^2\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[h_i\left(\hat{\phi}_i - \phi_i\right)^2\right]\right\} \text{ (Função de Verossimilhança)}$$

APÊNDICE 2 – Códigos de programação no software R

2.1 Geração das séries (simulação de dados)

2.1.1 Simulação dos valores de fi mediante Normal Truncada [-1,1]

```
#######pacotes exigidos##############
library(MASS)
library(MCMCpack)
library(gamlss.tr)
#######simulação das séries##########
m=20 #número de séries
n=7 #número de observação de cada série
e=matrix(0,m,(n+1)) #vetor de erros
y=matrix(0,m,(n+1)) #matriz de dados m x n
fi=matrix(0,m,1) #vetor de parâmetros m x 1
fi_trunc=trun.r(par=c(-1,1), family="NO", type="both") # gerar num. aleat. NORMAL truncada [-1,1]
fip=matrix(fi_trunc(m,mu = 0.5, sigma = 0.2),m,1) # fi paramétrico
sd<-sqrt(rinvgamma(1,shape=9,scale=10)) #gerar num. aleat. de gama inversa
 for(i in 1:m)
    {
      for(j in 1:(n+1))
       {
        e[i,j]<-rnorm(1,0,sd)

         for(j in 1:1)
          {y[i,j]<-e[i,j]}
            for(j in 2:(n+1))
            {y[i,j]<- fip[i]*y[i,j-1]  + e[i,j]}
       }
      }
ynobs=y[,(n+1)]#verdadeiro valor da obs yn+1
y=y[,1:n]
#######Funções auxiliares#############
h=matrix(0,m,1) #vetor de valores hi
fihat=matrix(0,m,1) #vetor de valores fi chapéu
for(i in 1:m)
    {
     h[i]=sum((y[i,(2:(n-1))])^2)
     fihat[i]=sum(y[i,2:n]*y[i,1:(n-1)])/h[i]
     }
fipar=t(fip)
fihat=t(fihat)
h=t(h)
```

OBS-1.: Para gerar valores de fi pelas distribuições Gumbel truncada (assimétrica) e uniforme (constante), utilizou-se respectivamente as linhas de comando 2.1.1 e 2.1.2.

2.1.2 Simulação dos valores de fi mediante Gumbel Truncada [-1,1]

```
fi_trunc=trun.r(par=c(-1,1), family="GU", type="both") #gerar num. aleat. de Gumbel truncada [-1,1]###
fip=matrix(fi_trunc(m,mu = 0.5, sigma = 0.35),m,1) # fi paramétrico
```

2.1.3 Simulação dos valores de fi mediante Uniforme [-1,1]

```
fip=matrix(runif(m, -1, 1),m,1) #gerar num. aleat. Uniforme [-1,1]###
```

2.2 Algoritmos MCMC: análise dos dados reais

```
#######pacotes exigidos##############
library(MASS)
library(MCMCpack)
library(gamlss.tr)
#######valores iniciais - máxima verossimilhança##########
dados=read.table("C:/Users/Fabyano/Desktop/fujita_fim1.txt",h=T)#leitura dos dados reais
a1=arima(dados$reckp[1:7] , order = c(1, 0, 0), method = c("ML"))
a2=arima(dados$cmycp[1:7] , order = c(1, 0, 0), method = c("ML"))
a3=arima(dados$srcp[1:7]  , order = c(1, 0, 0), method = c("ML"))
a4=arima(dados$timp2p[1:7], order = c(1, 0, 0), method = c("ML"))
a5=arima(dados$ikkap[1:7] , order = c(1, 0, 0), method = c("ML"))
a6=arima(dados$nfkbp[1:7] , order = c(1, 0, 0), method = c("ML"))
a7=arima(dados$nemop[1:7] , order = c(1, 0, 0), method = c("ML"))
a8=arima(dados$nikp[1:7]  , order = c(1, 0, 0), method = c("ML"))

fipar=rbind(a1$coef[1],a2$coef[1],a3$coef[1],a4$coef[1],a5$coef[1],a6$coef[1],a7$coef[1]
,a8$coef[1])
sd=sqrt(mean(rbind(a1$sigma2,a2$sigma2,a3$sigma2,a4$sigma2,a5$sigma2,a6$sigma2,a7$sigma2
,a8$sigma2)))
####### dimensionamento do conjunto de dados reais##########
m=8 #número de séries
n=7 #número de observação de cada série
y=t(dados[1:7,])
#######Funções auxiliares#############
h=matrix(0,m,1)#vetor de valores hi
fihat=matrix(0,m,1) #vetor de valores fi chapéu
for(i in 1:m)
    {
     h[i]=sum((y[i,(2:(n-1))])^2)
 fihat[i]=sum(y[i,2:n]*y[i,1:(n-1)])/h[i]
     }
fihat=t(fihat)
h=t(h)

######Gibbs e Metropolis##############
iter=20000
p=13.5
q=5 ## p e q devem ser especificado de acordo com os gráficos da dist. Beta (Fig. 3)
alfa=2
beta=3
 I1   = matrix(1,1,m)
 fi   = matrix(0,iter,m)
 cand = matrix(0,iter,m)
 sig2 = matrix(0,iter,1)

  for (k in 1:1)
   {
    for (i in 1:m)
     {
      fi[k,] = fipar
     }
      sig2[k] = sd^2
  }
```

```
  for (k in 2:iter)
   {
    for (i in 1:m)
    {
       fi[k,]     =  ((I1-fi[k-1,])^(q-0.5))*((I1+fi[k-1,])^(p-0.5))*exp(-0.5*(1/sig2[k-
                     1])*h*(fihat-fi[k-1,])^2) #condicional completa de fi
     cand[k,]  = mvrnorm(1,fipar,diag(0.000001,m,m))
         prob  = min(1, dnorm(mean(cand[k,]))/dnorm(mean(fi[k,])))
                 if (runif(1) < prob)
                 fi[k,] = cand[k,] else fi[k,] = fi[k-1,]
     }
       alfa1 = (n/2)*m + alfa*m + m -1
       beta1 = (sum((h*(fi[k]-fihat)^2))+2*beta)/2
     sig2[k] = rinvgamma(1,shape=alfa1,scale=beta1) #condicional completa de sigma2

    }
matfim=cbind(fi,sig2)
matfim1=matfim[(iter/2):iter,] # matriz de resultados

######distribuição preditiva##############

pred=matrix(0,iter,m)

 for(k in 1:nrow(matfim))
  {
   for(i in 1:m)
    {
     pred[k,]=mvrnorm(1,matfim[k,1:(ncol(matfim)-
              1)]*y[,ncol(y)],diag(matfim[k,ncol(matfim)],m,m))
    }
  }
matfim2=pred
matfim2=matfim2[(iter/2):iter,] # matriz de resultados

######fator de bayes###################

Si1=sum(dados$reckp[1:7])
Si2=sum(dados$cmycp[1:7])
Si3=sum(dados$srcp[1:7] )
Si4=sum(dados$timp2p[1:7])
Si5=sum(dados$ikkap[1:7])
Si6=sum(dados$nfkbp[1:7])
Si7=sum(dados$nemop[1:7])
Si8=sum(dados$nikp[1:7] )


i1=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,1]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si1,nrow(matfim1),1)+matrix(h[1],nrow(matfim1),1)*(matrix(fihat[
1]^2,nrow(matfim1),1) - 2*matfim1[,1] + matfim1[,1]^2)))

i2=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,2]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si2,nrow(matfim1),1)+matrix(h[2],nrow(matfim1),1)*(matrix(fihat[
2]^2,nrow(matfim1),1) - 2*matfim1[,2] + matfim1[,2]^2)))

i3=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,3]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si3,nrow(matfim1),1)+matrix(h[3],nrow(matfim1),1)*(matrix(fihat[
3]^2,nrow(matfim1),1) - 2*matfim1[,3] + matfim1[,3]^2)))
```

```
i4=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,4]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si4,nrow(matfim1),1)+matrix(h[4],nrow(matfim1),1)*(matrix(fihat[
4]^2,nrow(matfim1),1) - 2*matfim1[,4] + matfim1[,4]^2)))

i5=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,5]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si5,nrow(matfim1),1)+matrix(h[5],nrow(matfim1),1)*(matrix(fihat[
5]^2,nrow(matfim1),1) - 2*matfim1[,5] + matfim1[,5]^2)))

i6=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,6]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si6,nrow(matfim1),1)+matrix(h[6],nrow(matfim1),1)*(matrix(fihat[
6]^2,nrow(matfim1),1) - 2*matfim1[,6] + matfim1[,6]^2)))

i7=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,7]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si7,nrow(matfim1),1)+matrix(h[7],nrow(matfim1),1)*(matrix(fihat[
7]^2,nrow(matfim1),1) - 2*matfim1[,7] + matfim1[,7]^2)))

i8=(matfim1[,9])^(n/2)*(matrix(1,nrow(matfim1),1)-matfim1[,8]^2)*exp(-
0.5*matfim1[,9]*(matrix(Si8,nrow(matfim1),1)+matrix(h[8],nrow(matfim1),1)*(matrix(fihat[
8]^2,nrow(matfim1),1) - 2*matfim1[,8] + matfim1[,8]^2)))

prod=i1*i2*i3*i4*i5*i6*i7*i8

Ml_normal=sum(prod)/nrow(matfim1)  # estimativa da verossimilhança marginal via MCMC
```

OBS-2.: Os códigos no item anterior dizem respeito a implementação da priori simétrica, para as demais prioris adotadas, assimétrica e constante, os códigos foram os mesmo, porém foram alterados os valores de p e q (parâmetros da distribuição Beta). A comparação por meio do fator de Bayes foi realizada de acordo com a razão das verossimilhanças marginais (para a priori simétrica em questão, esta foi dada por Ml_normal).