MOYSÉS NASCIMENTO

ANÁLISE DE AGRUPAMENTO PARA DADOS EM PAINEL: APLICAÇÕES EM SÉRIES TEMPORAIS DE EXPRESSÃO GÊNICA

LAVRAS – MG
2011

MOYSÉS NASCIMENTO

ANÁLISE DE AGRUPAMENTO PARA DADOS EM PAINEL:
APLICAÇÕES EM SÉRIES TEMPORAIS DE EXPRESSÃO GÊNICA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientadora

Dr.ª Thelma Sáfadi

Coorientador

Dr. Fabyano Fonseca e Silva

LAVRAS – MG

2011

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Nascimento, Moysés.
Análise de agrupamento para dados em painel: aplicações em séries temporais de expressão gênica / Moysés Nascimento. – Lavras : UFLA, 2011.
121 p. : il.
Tese (doutorado) – Universidade Federal de Lavras, 2011.
Orientador: Thelma Sáfadi.
Bibliografia.

1. Microarray time series. 2. Modelo autorregressivo. 3. Predição da expressão gênica. 4. Séries temporais. I. Universidade Federal de Lavras. II. Título.

CDD – 519.536

MOYSÉS NASCIMENTO

ANÁLISE DE AGRUPAMENTO PARA DADOS EM PAINEL:
APLICAÇÕES EM SÉRIES TEMPORAIS DE EXPRESSÃO GÊNICA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Aprovada em 19 de abril de 2011.

Dr. Fabyano Fonseca e Silva	UFV
Dr. Luiz Alexandre Peternelli	UFV
Dr. Augusto Ramalho de Morais	UFLA
Dr. Daniel Furtado Ferreira	UFLA
Dr. Júlio Sílvio Souza Bueno Filho	UFLA

Dr.ª Thelma Sáfadi
(Orientadora)

LAVRAS - MG
2011

DEDICO

A minha avó,
Maria Neildes (in memoriam),
e minha esposa,
Ana Carolina.

AGRADECIMENTOS

Destaquei um agradecimento especial a algumas pessoas que tiveram um importante papel na minha formação, lembrando que estes não são todos.

Em primeiro lugar, gostaria de agradecer a Deus, que me preparou dando toda força e sabedoria necessária para os estudos e situações adversas deste desafio.

Gostaria também de destacar o trabalho especial que minha avó (*in memoriam*) e mãe tiveram na minha formação. Creio que elas tiveram o papel mais importante de toda a minha história. Fica minha gratidão pelo o apoio, seja na vida acadêmica, profissional ou pessoal.

A minha esposa e mulher da minha vida, Ana Carolina companheira, que nunca me abandonou, estando comigo, ora nos momentos difíceis, ora nas vitórias.

Aos meus orientadores Thelma Sáfadi e Fabyano Fonseca e Silva pela ajuda, exemplo e respeito dispensados ao longo da vida minha acadêmica. Obrigado pelo rico tesouro do conhecimento.

Aos professores (amigos) que atuaram na minha formação, em especial, Adésio Ferreira, Cosme Damião Cruz, Luiz Alexandre Peternelli e Mauro C. M. Campos que ajudaram a fortalecer o meu caráter e perfil profissional.

As amizades construídas em Lavras, em especial Augustão, Ana Paula, Ana Lúcia, Edcarlos, Leandro, Paulo e Tânia, pois estes participaram dos bons momentos vivenciados em Lavras.

A banca, composta pelos professores Fabyano Fonseca e Silva, Luiz Alexandre Peternelli, Augusto Ramalho de Morais, Daniel Furtado Ferreira e Júlio Sílvio Souza Bueno Filho que aceitaram o convite que lhes foi feito e, dessa forma, colaboraram para conclusão deste projeto.

À Universidade Federal de Lavras pela estrutura e oportunidade de desenvolver este projeto.

Ao CAPES pela concessão de um ano de bolsa para auxiliar no desenvolvimento deste projeto.

A todos os professores do Departamento de Estatística da Universidade Federal de Viçosa.

A todos os meus parentes e amigos, dos mais diversos meios. Vocês atuaram precisamente na minha vida, definindo exatamente o que sou.

RESUMO

Este trabalho teve por objetivo propor uma metodologia para o agrupamento de genes com padrões de expressões gênicas similares, baseado nas estimativas dos parâmetros provenientes do modelo autorregressivo de ordem p, AR(p), para dados em painel. Foram utilizados dados referentes à expressão de genes que atuam sobre ciclo celular de *Saccharomyces cerevisiae*, os quais correspondem a 114 genes, sendo que, cada um deles apresentava 10 valores de *fold-change* (medida da expressão) ao longo do tempo (0, 15, 30,...., 135 minutos). A estimação dos parâmetros do modelo AR(p) para dados em painel foi realizada sob dois diferentes enfoques. Sob o enfoque frequentista buscou-se verificar a viabilidade da utilização de métodos de agrupamentos, hierárquico (Ward) e de otimização (Tocher), na formação de grupos homogêneos de séries de expressão gênica para posterior ajuste de modelos autorregressivos, AR(2), para dados em painel. Os resultados obtidos por este enfoque indicaram que o método de Ward mostrou-se mais apropriado para a obtenção de grupos homogêneos, e a eficiência de predição da expressão gênica resultante do ajuste do modelo AR(2) para dados em painel foi de 100%. Sob o enfoque bayesiano, considerou-se de forma conjunta a análise de agrupamento por meio do método de Ward e a análise bayesiana do modelo AR(p) para dados em painel. Os resultados obtidos por este enfoque mostraram que a metodologia proposta foi capaz de agrupar genes que apresentavam padrões de expressão similares e também de proporcionar previsões eficazes para valores futuros da expressão gênica, obtidos por meio da teoria de distribuições preditivas.

Palavras-chave: *Microarray time series.* Previsão. *Saccharomyces cerevisiae.*

ABSTRACT

This work has as objective to propose a methodology for clustering genes with similar expression patterns based on parameter estimates from the autoregressive model of order p, AR (p), for panel data. Were used data from genes expression that are related with the *Saccharomyces cerevisiae* cell cycle. These data correspond to 114 genes, which each one had 10 fold-change values (expression measure) over time (0, 15, 30, ..., 135 minutes). The parameter estimation of AR(p) panel data model was realized by two different approaches. Under a frequentist approach, we aimed to verify the efficiency of hierarchical (Ward) and optimization (Tocher) clustering methods in the obtaining homogeneous clusters in order to fit the AR(2) panel data model to gene expression series. The results obtained under this approach indicated that the Ward method was more appropriate to obtain homogeneous cluster in relation to gene expression pattern. Furthermore, the efficiency of the gene expressions forecasting from AR(2) panel data model were 100%. Under a bayesian approach, was considered a joint study of cluster (Ward method) and bayesian analyses of AR (p) model for panel data. The results obtained by this approaches showed that the used methodology provided clustering with similar gene expressions pattern and also effective forecasting for the expression values in future times by predictive distribution theory.

Key words: Microarray time series. Forecasting. *Saccharomyces cerevisiae.*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

1.1 Contextualização

A estrutura de dados em painel, que representa um conjunto de dados longitudinais, é caracterizada pela combinação de várias séries temporais provenientes de diferentes unidades amostrais, as quais podem ser definidas, por exemplo, como diferentes tratamentos, indivíduos ou classes.

Sob o enfoque de séries de temporais, a utilização de dados em painel proporciona ao pesquisador um aumento na precisão das estimativas dos parâmetros de interesse em relação a análises individuais de cada série. Essa maior precisão é atribuída ao aumento no número de observações devido à combinação dos vários períodos de tempo para cada indivíduo avaliado.

Estudos envolvendo estrutura de dados em painel foram desenvolvidos em diversas áreas do conhecimento, tais como Econometria, no qual se avaliou a dinâmica temporal da criminalidade nos estados brasileiros (SANTOS, 2009), e em Medicina, avaliando aspectos sociais de políticas de saúde pública (SOUZA; LEITE FILHO, 2008). Além dessas, uma área que merece destaque, é a de Genética e Melhoramento, na qual foram realizados estudos voltados para a previsão de valores genéticos (SILVA et al., 2008b; SILVA et al., 2008a) e de expressão gênica (MORAIS et al., 2010) em instantes de tempo não observados. Esta última aplicação caracteriza-se como um grande avanço tecnológico, tendo em vista o alto custo proveniente de rotinas laboratoriais na determinação da expressão gênica e a grande demanda de tempo e mão de obra empregada.

Dentre os diversos modelos utilizados em estudos de dados em painel sob o ponto de vista de séries temporais, destaca-se o modelo autorregressivo

(AR), pois o mesmo é aplicado a diversas situações práticas e, em geral, apresenta boa qualidade de ajuste quando comparado com modelos mais complexos. Segundo Balgobin e Petruccelli (1997), a utilização de modelos autorregressivos para o ajuste de dados em painel não apenas produz previsões acuradas de valores futuros, mas também permite a análise de séries medidas em poucos instantes de tempo.

Independentemente da metodologia utilizada, bayesiana ou frequentista, a análise de modelos AR para dados em painel considera a combinação de informações de todos os indivíduos na estimação dos coeficientes individuais de cada série, de forma que estes coeficientes são considerados amostras aleatórias de uma mesma distribuição. Desse modo, faz-se necessário a utilização de modelos hierárquicos, os quais muitas vezes tornam a modelagem bastante complexa (LIU; TIAO, 1980). Sob o ponto de vista da inferência bayesiana, esta hierarquia pode ser incorporada, devido à possibilidade de se utilizar distribuições a priori para os parâmetros a serem estimados.

Segundo Hsiao e Sun (2000) a análise de dados em painel para grupos de séries não homogêneas pode gerar dificuldades para a especificação dos modelos e para a estimação dos parâmetros. Tendo em vista este problema, Silva et al. (2008b) utilizaram um procedimento empírico de agrupamento com o objetivo de homogeneizar o conjunto de séries temporais, caracterizando assim grupos homogêneos requisitados pela teoria dos paineis. Por outro lado, a utilização de métodos impessoais, tais como agrupamentos hierárquicos e de otimização, podem melhorar expressivamente a obtenção de tais grupos.

No campo da Estatística Genética, um tema que tem atraído interesse de pesquisadores é a análise de dados de expressão gênica identificada ao longo do tempo, os quais são denominados *Microarray Time Series* (*MTS*) (MUKHOPADHYAY; CHATTERJEE, 2007). O estudo destes dados tem possibilitado o entendimento de diversos processos biológicos, porém, devido à

grande quantidade de genes envolvidos e o pequeno número de medidas ao longo do tempo, e dado o alto custo dos processos laboratoriais, a análise de *MTS* se tornou um grande desafio (ERNST; NAU; BAR-JOSEPH, 2005). Assim, encontrar grupos de genes (séries de expressão gênica) que se expressem de forma similar possibilita aos pesquisadores inferir sobre a função e mecanismos de regulação gênica (COSTA; CARVALHO; SOUTO, 2004). Com esse objetivo trabalhos na literatura são dedicados a utilização de métodos de análise de agrupamento para a obtenção de grupos homogêneos de genes.

Dentre os métodos de análise de agrupamento utilizados, se destaca os métodos hierárquicos (EISEN et al., 1998). Entretanto, mesmo estes métodos sendo extremamente utilizados em problemas biológicos, os mesmos não são desenvolvidos para utilização em dados de séries temporais, pois ignoram a natureza sequencial das observações. Para contornar esse problema, inerente aos métodos de agrupamentos hierárquicos, Ramoni, Sebastiani e Kohane (2002) agruparam os genes baseando-se na dinâmica do padrão da expressão, por meio da escolha de um modelo M_c, para um conjunto de c grupos de séries temporais, constituído de c modelos autorregressivos, que maximiza a sua probabilidade a *posteriori*. Além desse, podem-se citar os trabalhos de Schliep, Schonhuth e Steinhoff (2003), que diferentemente do trabalho de Ramoni, Sebastiani e Kohane (2002), fizeram uso de modelos de Markov. Ocultos para obtenção de uma solução para o problema de agrupamento de genes e, Bar-Joseph et al. (2003) que realizaram o agrupamento dos genes por meio de um modelo de misturas de representações contínuas do tipo B-splines da expressão gênica temporal. Entretanto, apesar destes métodos serem úteis, segundo Bar-Joseph (2004) os mesmos não funcionam bem para experimentos relativamente pequenos, isto é, experimentos que possuem menos de 10 observações temporais.

Segundo Ernst, Nau e Bar-Joseph (2005) mais de 80% de todas as séries contidas no banco de dados de Stanford (*Stanford Microarray Database* - *SMD*) possuíam menos de 8 observações temporais. Essa característica da análise de *MTS* juntamente com a grande quantidade de genes avaliados neste tipo de estudo, nos leva à busca de alternativas mais eficientes para o agrupamento dos genes com padrões de expressão similares.

De acordo com o exposto pode-se perceber que as características inerentes a análise de *MTS* estão diretamente ligadas à utilização de uma estrutura de dados em painel. Desta forma, acredita-se que uma nova metodologia de agrupamento dos padrões de expressão gênica, tomando-se por base estimativas dos parâmetros provenientes de uma análise de dados em estrutura de painel, minimizaria o problema referente ao pequeno número de observações temporais e proporcionaria maior acurácia na solução do problema de agrupamento.

1.2 Objetivos do trabalho

Este trabalho teve como objetivo geral propor uma metodologia para o agrupamento de genes com padrões de expressões gênicas similares, baseado nas estimativas dos parâmetros provenientes da análise bayesiana do modelo autorregressivo de ordem p, AR(p), para dados em painel.

Dentre os objetivos específicos, destacam-se:

1. Verificar qual a melhor alternativa, dentre os métodos de agrupamentos hierárquicos (Ward) e de otimização (Tocher), na formação de grupos homogêneos de séries de expressão gênica para posterior ajuste de modelos autorregressivos, AR(p), para dados em painel;

2. apresentar a base teórica para a análise bayesiana do modelo autorregressivo para dados em painel e para o novo método de agrupamento proposto;
3. fornecer um algoritmo baseado no método *MCMC* (*Markov Chain Monte Carlo*) capaz de estimar os parâmetros do modelo bayesiano AR para dados em painel e agrupar as séries temporais de expressão gênica por meio das estimativas obtidas. Sendo que, ao final do processo seja fornecido o número de grupos e de suas partições automaticamente.

1.3 Organização do trabalho

Esta tese está organizada em formato de artigo, de acordo com as normas da Universidade Federal de Lavras - UFLA (2011):

1. Na primeira parte apresenta-se uma introdução na qual o trabalho é contextualizado e seus objetivos são descritos. A segunda parte desta tese é composta por dois artigos.

1.1 O artigo 1 teve por objetivo verificar a viabilidade da utilização de métodos de agrupamentos, hierárquico (Ward) e de otimização (Tocher), na formação de grupos homogêneos de séries de expressão gênica para posterior ajuste de modelos autorregressivos, AR(p), para dados em painel. Desta forma tal ajuste possibilita a realização de previsões da expressão gênica dentro de cada grupo formado.

1.2 O artigo 2 apresenta uma proposta metodológica para o agrupamento de genes com padrões de expressões gênicas similares, baseadas nas estimativas dos parâmetros provenientes da análise bayesiana do modelo autorregressivo de ordem p, AR(p), para dados em painel. Além disso, propõe-se

também realizar previsões baseadas em distribuições preditivas para valores de expressões gênicas em tempos futuros.

2. A última parte do trabalho é composta por uma seção denominada de considerações finais em que são apresentados, de forma sucinta, todos os aspectos do trabalho e algumas perspectivas de estudos futuros.

2 REFERENCIAL TEÓRICO

2.1 Dados em painel

Uma estrutura de dados em painel consiste da combinação de várias séries temporais provenientes de diferentes unidades amostrais, as quais podem ser definidas, por exemplo, como diferentes tratamentos, indivíduos ou classes.

De acordo com Arellano e Bover (1990), a utilização de dados em painel, para o estudo de diversos fenômenos de interesse, tornou-se cada vez mais frenquente na literatura a partir da década de 80. Este crescimento se deve ao aumento da disponibilidade deste tipo de estrutura de dados. Esta estrutura, que representa um conjunto de dados longitudinais, é caracterizada pela combinação de várias séries temporais provenientes de diferentes unidades amostrais, as quais podem ser definidas, por exemplo, como diferentes tratamentos, indivíduos ou classes.

Segundo Cameron e Trivedi (2005) a análise de dados em painel tem como sua maior vantagem o aumento da precisão na estimação dos parâmetros. Este ganho de precisão é resultante do aumento do número de observações devido à combinação de vários períodos de tempo para cada indivíduo.

A utilização de dados em painel para a análise de dados vem sendo considerada por profissionais em diversas áreas de conhecimento, tais como o trabalho de Souza e Leite Filho (2008), em que os autores analisaram os fatores determinantes do status de saúde em cada estado da Região Nordeste do Brasil. Rivero et al. (2009) os quais utilizaram modelos de regressão linear com dados em painel para analisar a evolução das causas imediatas do desmatamento da Amazônia. Sob o enfoque de série temporal, Silva et al. (2008b) realizaram uma análise bayesiana do modelo autorregressivo aplicado para dados em painel para previsão de valores genéticos de touros da raça Nelore. A análise bayesiana para

dados em painel foi também empregada por Morais et al. (2010) na modelagem da expressão gênica para oito genes.

Na Tabela 1 está apresentada a organização de um conjunto de dados em painel, em que Y_{ij} é valor do i-ésimo indivíduo no tempo j.

Tabela 1 Representação teórica de um conjunto de dados organizados em estrutura de painel, para m indivíduos avaliados em n_i tempos

Indivíduo	Tempo					
	1	2	...	t	...	n_i
1	Y_{11}	Y_{12}	...	Y_{1t}	...	Y_{1n_1}
2	Y_{21}	Y_{22}	...	Y_{2t}	...	Y_{1n_2}
.	.	.	...		...	.
.	.	.	...		...	.
.	.	.	...		...	.
m	Y_{m1}	Y_{m2}	...	Y_{mt}	...	Y_{1n_m}

2.1.2 Modelos autorregressivos para dados em painel

A literatura a respeito da modelagem e estimação para dados em painel é bastante vasta e complexa. A modelagem pode ser realizada por meio de modelos de regressão linear e modelos de séries temporais.

Dentre os modelos de regressão, para o ajuste de dados em painel, o mais simples é o conhecido como "*pooled*". Este modelo considera que todos os coeficientes são constantes, isto é, os mesmos não variam de acordo com o tempo e nem conforme os indivíduos. Uma variante do modelo "*pooled*" é

definida quando consideram-se que os interceptos variam entre os indivíduos (CAMERON; TRIVEDI, 2005).

Além dos modelos de regressão o ajuste para dados em painel pode ser realizado por modelos de séries temporais. Dentre os diversos modelos aplicados em análise de séries temporais, um dos mais estudados e utilizados é o modelo autorregressivo de ordem p, denotado por AR(p).

Sob o enfoque bayesiano, a modelagem do modelo AR(p), é tema de grande interesse para alguns pesquisadores. Liu e Tiao (1980), Nandram e Petruccelli (1997) e Silva et al. (2011) apresentaram a análise bayesiana para o modelo AR(p) quando se dispõe de uma estrutura de dados em painel.

Quando se possui mais de uma série em estudo, pode-se generalizar o modelo AR (p) através da utilização dos vetores autorregressivos (VAR – Vector Autoregressive) (HOLTZ-EAKIN; NEWEY; ROSEN, 1988).

A respeito da análise de *Microarray Time Séries* (*MTS*), a utilização de modelos autorregressivos no estudo da expressão gênica pode ser encontrada em alguns trabalhos disponíveis na literatura. Dentre estes, pode-se citar o trabalho de Fujita et al. (2007a) em que foram utilizados vetores autorregressivos (*VAR – Vector Autoregressive*) para descrever o comportamento de índices de expressão gênica ao longo de diferentes horas. Fujita et al. (2007b) descreveram o método de vetores autorregressivos (VAR) para busca de relações de causalidade entre os genes, analisando os níveis de expressão do gene sem prévio conhecimento biológico. Morais et al. (2010) modelaram a expressão gênica temporal, para um pequeno número de genes, por meio da análise bayesiana de um modelo AR(1) para dados em painel.

No trabalho desenvolvido por Morais et al. (2010), devido à pequena quantidade de genes em estudo, assumiu-se a homogeneidade requerida para análise de dados em painel (LIU; TIAO, 1980). Entretanto, devido à grande quantidade de genes envolvidos numa análise de *MTS*, essa pressuposição de

homogeneidade não pode ser sempre assumida verdadeira. Assim, faz-se necessário, e será abordada nesse trabalho, a utilização de um procedimento para a obtenção de grupos de genes homogêneos, de forma a possibilitar a utilização do modelo autorregressivo para dados em painel.

O modelo autorregressivo de ordem p, AR(p), para dados em painel, em que p é o número de parâmetros do modelo, é representado pela equação 1 (SILVA et al., 2011):

$$Y_{it} = \mu_i + \phi_{i1} Y_{i(t-1)} + \phi_{i2} Y_{i(t-2)} + ... + \phi_{ip} Y_{i(t-p)} + e_{it}$$

ou

$$Y_{it} = \mu_i + \sum_{j=1}^{p} \phi_{ij} Y_{i(t-j)} + e_{it} \qquad (1)$$

em que: i=1,2, ...,m; j=1,2,...,p; t=1,2,..., n_i; μ_i é a média do processo referente ao individuo i; Y_{it} é o valor atual de um processo estocástico referente ao indivíduo i, $\phi_{i1}, \phi_{i2}, \ldots, \phi_{ip}$ são os parâmetros referentes ao modelo para o i-ésimo indivíduo, denominados parâmetros de autorregressão; e_{it} é o resíduo associado ao modelo, também denominado de ruído branco, $e_{it} \overset{iid}{\sim} N(0, \sigma_e^2)$.

De acordo com esta notação, tem-se m indivíduos, com n_i observações longitudinais cada, indicando que os indivíduos podem apresentar números diferentes de observações. Além disso, o modelo contempla p parâmetros por indivíduo.

A equação 1, pode ser reescrita da seguinte forma:

$$\left(1 - \phi_{i1} B - \phi_{i2} B^2 - \ldots - \phi_{ip} B^p\right) Y_{it} = e_{it} \qquad (2)$$

em que:; $B^s Y_{it}$ é um operador translação para o passado, ou seja, $B^s Y_{it} = Y_{i(t-s)}$ para s=1,2,...p. De acordo com Morettin e Toloi (2008), Y_{it} é considerado um processo estacionário se as raízes de $\phi_i(B) = \left(1 - \phi_{i1}B - \phi_{i2}B^2 + \ldots + \phi_{ip}B^p\right) = 0$ estiverem fora do circulo unitário. Para modelos AR(1) e AR(2) essa condição é satisfeita, para cada série individual, respectivamente por $|\phi_{i1}| < 1$ e $|\phi_{i2} - \phi_{i1}| < 1$, $|\phi_{i1} + \phi_{i2}| < 1$, $-1 < \phi_{i2} < 1$.

2.2 Inferência bayesiana

Estatística é uma área de conhecimento que lida com problemas nos quais quantidades aleatórias estão envolvidas. Particularmente, na inferência Estatística o interesse recai numa quantidade desconhecida e não observada (θ), em que θ assume valores no conjunto Θ. Essa quantidade pode ser um escalar, um vetor, ou mesmo uma matriz. O principal problema da área consiste em descrever a incerteza sobre θ (PAULINO; TURKMAN; MURTEIRA, 2003).

Na inferência clássica θ é apenas um parâmetro desconhecido e a única fonte de informação relevante sobre este parâmetro é a informação probabilística de quantidades aleatórias “observáveis” associadas a ele. Por outro lado, na inferência bayesiana, a abordagem é um pouco diferente. A diferença essencial é que θ é pensado como uma quantidade aleatória, tal como os “observáveis" associados a ele, e assim outras fontes de informação são consideradas.

Denote por H a informação inicial disponível sobre θ. Assuma que essa informação possa ser expressa em termos probabilísticos por meio de uma distribuição de probabilidade em θ, genericamente denotada por $p(\theta \mid H)$. Se a informação contida em H é suficiente, então a descrição da incerteza sobre θ está completa (PAEZ; GAMERMAN, 2005).

Entretanto, na maioria dos casos a informação inicial H não é suficiente. Nesse caso, a informação inicial precisa ser aumentada e a principal ferramenta utilizada nessa tarefa é a experimentação. Assuma que um vetor $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de quantidades aleatórias relacionadas à θ possa ser observado. Este vetor proporciona informação adicional sobre θ. Assume-se também que a distribuição amostral de X dado θ e H, denotada por $f(x \mid \theta, H)$, é conhecida.

Desta forma, a informação sobre θ está resumida pela distribuição $p(\theta \mid x, H)$. Utilizando o teorema de Bayes[1] pode-se relacionar $p(\theta \mid x, H)$ com $p(\theta \mid H)$ e $f(x \mid \theta, H)$. De fato

$$p(\theta \mid x, H) = \frac{f(\theta, x \mid H)}{f(x \mid H)} = \frac{f(x \mid \theta, H)p(\theta \mid H)}{\int_{\Theta} f(x \mid \theta, H)p(\theta \mid H)d\theta}$$

Para simplificar a notação, vamos omitir a dependência em H visto que ela aparece em todos os termos. Além disso, observa-se que a função no denominador não depende de θ, portanto é só uma constante em relação a $p(\theta \mid x)$. Assim, pode-se reescrever o teorema de Bayes da seguinte maneira

$$p(\theta \mid x,) = k \times f(x \mid \theta)p(\theta) \propto f(x \mid \theta)p(\theta)$$

A equação anterior proporciona uma regra para atualizar probabilidades sobre θ, partindo de $p(\theta)$ e chegando a $p(\theta \mid x)$. Daí a razão para chamar $p(\theta)$ de distribuição a *priori* e $p(\theta \mid x)$ de distribuição a *posteriori*. A função $f(x \mid \theta)$

---

[1] Teorema de Bayes: Suponha que eventos $C_1, C_2, \ldots, C_k$ formem uma partição de Ω e que suas probabilidades sejam conhecidas. Suponha ainda que para um evento A, se conheçam as probabilidades $P(A \mid C_i)$ para todo i=1,2,...,k. Então para qualquer j,

$$P(C_j \mid A) = \frac{P(A \mid C_j)P(C_j)}{\sum_{i=1}^{k} P(A \mid C_i)P(C_i)}.$$

é conhecida como função de verossimilhança de θ correspondente à amostra observada **X** = **x** (PAEZ; GAMERMAN, 2005).

A distribuição a *posteriori* descreve completamente a incerteza sobre θ após a observação dos dados, levando em conta a distribuição a *priori*. Isso representa uma distinção importante entre a inferência clássica e a bayesiana, visto que na abordagem clássica a incerteza sobre θ é descrita via o cálculo exato ou estimação (o que é mais comum) do erro padrão de um estimador pontual proposto de forma criteriosa para θ. Outra observação é que a distribuição a *posteriori* depende dos dados somente através de $f(x \mid \theta)$.

A integração da distribuição conjunta a *posteriori* para a obtenção das marginais, no caso multiparamétrico, geralmente não é analítica, necessitando de algoritmos iterativos especializados como o *Gibbs Sampler* (Amostrador de Gibbs) e/ou Metropolis-Hastings, os quais são denominados de algoritmos *MCMC* (*Markov Chain Monte Carlo*).

2.3 Métodos Monte Carlo via cadeias de Markov (*MCMC*)

Desde a década de 90, principalmente devido aos avanços dos recursos computacionais, os métodos de simulação *MCMC* passaram a ser tema obrigatório para profissionais em diversas áreas. Estes métodos surgiram como alternativa para solução de problemas complexos em inferência estatística (clássica e bayesiana).

Os métodos *MCMC* têm como proposta simular da distribuição de interesse (π) via construção de uma cadeia de Markov em seu suporte tendo π como sua única distribuição estacionária. Os métodos *MCMC* garantem que, após um tempo suficientemente longo de simulação, elementos de Λ podem ser amostrados com distribuição aproximadamente igual a π. O conjunto Λ é

chamado de espaço de estados e cada um de seus elementos é chamado de estado.

2.3.1 Algoritmo de Metropolis-Hastings

O primeiro algoritmo *MCMC* conhecido por algoritmo de Metropolis é, sem dúvida, o mais fundamental dos métodos *MCMC*, pois todos os outros são derivações dele. O algoritmo originalmente foi proposto por Metropolis et al. (1953) e generalizado por Hastings (1970).

Suponha que $X = (X_1, X_2, \ldots, X_d)$, seja um vetor aleatório discreto d-dimensional com distribuição de probabilidade π, cujo espaço amostral é um conjunto finito Λ. É importante ressalvar que a exposição restrita ao caso em que Λ é finito não perde em generalidade.

O algoritmo de Metropolis-Hastings propõe uma cadeia de Markov $(X_n)_{n\geq 0}$ em Λ com distribuição estacionária π. Os ingredientes básicos são:

- Uma função de transição auxiliar $q(x, y)$ tal que:

- $0 \leq q(x, y) \leq 1$, para todo $(x, y) \in \Lambda \times \Lambda$;

- $\sum_{y \in \Lambda} q(x, y) = 1$, para todo $x \in \Lambda$.

- Uma função $\alpha(x, y)$ tal que:

- $0 \leq \alpha(x, y) \leq 1$, para todo $(x, y) \in \Lambda \times \Lambda$;

- $\alpha(x, x) = 1$, para todo $x \in \Lambda$.

A ideia básica destes métodos é fazer com que a função de transição dos algoritmos possua a condição de reversibilidade. Essa condição garante que se π é uma distribuição reversível para a cadeia, então a mesma é também uma distribuição estacionária e única para a cadeia (CHIB; GREENBERG, 1995).

Para que essa condição seja satisfeita, basta que a probabilidade de mudança de um estado x para o estado y seja definida como:

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$$

De posse dos ingredientes necessários o algoritmo de Metropolis-Hastings pode ser descrito da seguinte forma:

1. Escolha uma função de transição $q(x, y)$;
2. Escolha $X_0 \in \Lambda$;
3. Para $n \geq 0$ e $X_n = x$ simule uma realização de $Y \sim q(x, y)$ e lance uma $u \sim U(0,1)$. Supondo que $Y = y$, faça

$$X_{n+1} = \begin{cases} y & \text{se } u < \alpha(x, y); \\ x & \text{caso contrário.} \end{cases}$$

4. $n \leftarrow n + 1$ e retorne para o passo 3 até obter a convergência.

2.3.2 Amostrador de Gibbs

Outro importante método *MCMC* é o Amostrador de Gibbs (Gibbs *sampler*), este é simplesmente um caso particular do algoritmo de Metropolis-Hastings. A discussão exige a seguinte notação:

- $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$;
- $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$;
- $\pi_i(x_i \mid \mathbf{x}_{-i}) = P(X_i = x_i \mid \mathbf{X}_{-i} = \mathbf{x}_{-i})$.

As distribuições $\pi_i(\cdot \mid \mathbf{x}_{-i})$ são conhecidas como distribuições condicionais completas. Uma característica que torna o método interessante é que apenas essas distribuições são utilizadas na simulação. Assim, mesmo

desejando simular valores de uma distribuição de alta dimensão, as simulações são feitas através de uma única distribuição (CASELLA; GEORGE, 1992).

A cadeia do algoritmo é definida da seguinte maneira: Se $\mathbf{X}_n = \mathbf{x} = (x_1, \ldots, x_d)$ escolha uniformemente um índice em $\{1,2,\ldots,d\}$. Se o índice escolhido foi i, então simule um valor $X \sim \pi_i(x_i \mid \mathbf{x}_{-i})$. Se X=x, então o vetor candidato é dado por $\mathbf{x} = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots x_d)$ em que $q(y \mid x) = \frac{\pi_i(x_i \mid \mathbf{x}_{-i})}{d}$.

Quando utiliza-se o amostrador de Gibbs, o vetor candidato é sempre aceito como o próximo estado da cadeia.

É possível atualizar componente a componente do estado da cadeia do algoritmo de forma sequencial. Assim, o amostrador de Gibbs pode ser descrito tal como segue:

1. Escolha $\mathbf{X}_0 = (X_{10}, \ldots, X_{d0})$ em Λ;

2. Para $n \geq 0$, obtenha o estado $\mathbf{X}_{n+1} = (X_{1n+1}, \ldots, X_{dn+1})$ de $\mathbf{X}_n = (X_{1n}, \ldots, X_{dn})$ via simulação sequencial dos valores

$$X_{1n+1} \sim \pi_1(x_1 \mid x_{2n}, x_{3n}, \ldots, x_{dn})$$

$$X_{2n+1} \sim \pi_2(x_2 \mid x_{1n}, x_{3n}, \ldots, x_{dn})$$

$$\vdots$$

$$X_{dn+1} \sim \pi_d(x_2 \mid x_{1n}, x_{2n}, \ldots, x_{d-1n})$$

3. Faça $n \leftarrow n+1$ e retorne a passo 2, até obter a convergência.

Os algoritmos MCMC são processos iterativos, portanto surgem questões referentes à avaliação de suas convergências, ou seja, como verificar se os valores gerados seguem uma distribuição aproximadamente igual à distribuição de interesse. Na literatura são apresentados vários métodos necessários para a realização desta avaliação, e dentre estes se destacam Geweke

(1992), Heidelberger e Welch (1993) e Raftery e Lewis (1992). Uma maneira prática de aplicar todos estes métodos é por meio da biblioteca BOA - *Bayesian Output Analysis Program* (SMITH, 2007) do *software* livre R (R DEVELOPMENT CORE TEAM, 2008).

2.4 Análise bayesiana do modelo autorregressivo para dados em painel

O modelo autorregressivo de ordem p, AR(p), para dados em painel, em que p é o número de parâmetros do modelo, é representado pela equação 1 (SILVA et al., 2011):

$$Y_{it} = \mu_i + \phi_{i1} Y_{i(t-1)} + \phi_{i2} Y_{i(t-2)} + ... + \phi_{ip} Y_{i(t-p)} + e_{it}$$

ou

$$Y_{it} = \mu_i + \sum_{j=1}^{p} \phi_{ij} Y_{i(t-j)} + e_{it} \qquad (1)$$

em que: i=1,2, ...,m; j=1,2, ...,p; t=1,2,..., n_i; μ_i é a média de um processo estacionário referente ao indivíduo i; Y_{it} é o valor atual do processo, referente ao indivíduo i, $\phi_{i1}, \phi_{i2}, \ldots, \phi_{ip}$ são os parâmetros referente a equação 1 para o i-ésimo indivíduo, denominados parâmetros de auto-regressão; e_{it} é o resíduo associado ao modelo, também denominado de ruído branco, $e_{it} \overset{iid}{\sim} N(0, \sigma_e^2)$.

A função de verossimilhança, considerando $n_1 = n_2 = \ldots = n_m = n$, de acordo como o modelo apresentado em (1) condicionado nas p primeiras observações de cada indivíduo, é dada por:

$$L\left(\mathbf{Y} \mid \mathbf{\Phi}, \sigma_e^2\right) \propto \sigma_e^{2\,-\left(\frac{m(n-p)}{2}\right)} \exp\left\{ -\frac{1}{2\sigma_e^2} \sum_{i=1}^{m} \sum_{t=p+1}^{n} \left(Y_{it} - \mu_i - \sum_{j=1}^{p} \phi_{ij} Y_{i(t-j)} \right)^2 \right\}.$$

Reescrevendo a função de verossimilhança em forma matricial para todos os indivíduos, tem-se:

$$L\left(\mathbf{Y} \mid \mathbf{\Phi}, \sigma_e^2\right) \propto \sigma_e^{2\,\frac{-m(n-p)}{2}} \exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{Y}-\mathbf{X\Phi}\right)'\left(\mathbf{Y}-\mathbf{X\Phi}\right)\right\},$$

em que: $\mathbf{Y} = \left[y_{1(p+1)}, y_{1(p+2)}, \ldots, y_{1(n)}, y_{2(p+1)}, \ldots, y_{2(n)}, \ldots, y_{m(p+1)}, \ldots, y_{m(n)}\right]'$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & 0 & 0 \\ 0 & \mathbf{X}_2 & 0 & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{X}_m \end{bmatrix}_{m(n-p)\times m(p+1)},$$

$$\mathbf{X}_i = \begin{bmatrix} 1 & y_{i(p)} & \cdots & y_{i(1)} \\ 1 & y_{i(p+1)} & \cdots & y_{i(2)} \\ \vdots & & \ddots & \vdots \\ 1 & y_{i(n-1)} & \cdots & y_{i(n-p)} \end{bmatrix}_{(n-p)\times(p+1)} \quad \text{e}$$

$$\mathbf{\Phi} = \left[\mu_1, \phi_{11}, \phi_{12}, \ldots, \phi_{1p}, \mu_1, \phi_{21}, \ldots, \phi_{2p}, \ldots, \mu_m, \phi_{m1}, \ldots, \phi_{mp}\right]' \in R^{m(p+1)}.$$

Os vetores $\mathbf{\Phi}$ e $\mathbf{Y}$, considerados nas expressões apresentam, respectivamente, as dimensões m(p+1) x 1 e m(n-p) x 1 (SILVA et al., 2008b).

De acordo com a metodologia bayesiana, para a estimação dos parâmetros do modelo AR(p), faz-se necessário atribuir distribuições a *priori* para os parâmetros de interesse $\mathbf{\Phi}$ e σ_e^2. Neste estudo, da mesma forma que no trabalho de Silva et al. (2008b), considerou-se a *priori* hierárquica Normal multivariada - Gama Inversa, representada como segue:

$$P\left(\mathbf{\Phi}, \sigma_e^2\right) = P\left(\mathbf{\Phi} \mid \sigma_e^2\right) P\left(\sigma_e^2\right)$$

em que: $\left(\mathbf{\Phi} \mid \sigma_e^2\right) \sim N_{m(p+1)}\left(\boldsymbol{\mu}, \sigma_e^2 \mathbf{I}\right)$ e $\sigma_e^2 \sim GI(\alpha, \beta)$ (Gama Inversa), em que $\mathbf{I}$ é uma matriz identidade de dimensões m(p+1) x m(p+1). Assim, tem-se:

$$P\left(\boldsymbol{\Phi} \mid \sigma_e^2\right) \propto \sigma_e^{2^{-\left(\frac{m(p+1)}{2}\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[(\boldsymbol{\Phi}-\boldsymbol{\mu})' \mathbf{I}(\boldsymbol{\Phi}-\boldsymbol{\mu})\right]\right\} \text{ e}$$

$$P\left(\sigma_e^2\right) \propto \sigma_e^{2^{-(\alpha+1)}} \exp\left\{-\frac{\beta}{\sigma_e^2}\right\}.$$

Portanto, a distribuição conjunta a *priori*, $P\left(\boldsymbol{\Phi}, \sigma_e^2\right)$, é dada por:

$$P\left(\boldsymbol{\Phi}, \sigma_e^2\right) \propto \sigma_e^{2^{-\left(\frac{m(p+1)+2\alpha}{2}+1\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[2\beta + (\boldsymbol{\Phi}-\boldsymbol{\mu})' \mathbf{I}(\boldsymbol{\Phi}-\boldsymbol{\mu})\right]\right\}.$$

Os componentes $\boldsymbol{\mu}$, $\mathbf{I}$ (matriz identidade), α e β são denominados hiperparâmetros, e representam os parâmetros das distribuições a *priori* dos parâmetros do modelo considerado.

Combinando a função de verossimilhança, $L\left(\mathbf{Y} \mid \boldsymbol{\Phi}, \sigma_e^2\right)$, com a distribuição a *priori*, $P\left(\boldsymbol{\Phi}, \sigma_e^2\right)$, obtém-se, via Teorema de Bayes, a distribuição conjunta a *posteriori*:

$$P\left(\boldsymbol{\Phi}, \sigma_e^2 \mid \mathbf{Y}\right) \propto L\left(\mathbf{Y} \mid \boldsymbol{\Phi}, \sigma_e^2\right) P\left(\boldsymbol{\Phi}, \sigma_e^2\right),$$

$$P\left(\boldsymbol{\Phi}, \sigma_e^2 \mid \mathbf{Y}\right) \propto \sigma_e^{2^{\frac{-m(n-p)}{2}}} \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\Phi})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\Phi})\right\} \times$$

$$\sigma_e^{2^{-\left(\frac{m(p+1)+2\alpha}{2}+1\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[2\beta + (\boldsymbol{\Phi}-\boldsymbol{\mu})' \mathbf{I}(\boldsymbol{\Phi}-\boldsymbol{\mu})\right]\right\},$$

$$P\left(\boldsymbol{\Phi}, \sigma_e^2 \mid \mathbf{Y}\right) \propto \sigma_e^{2^{-\left(\frac{m(n+1)+2\alpha}{2}+1\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[2D + \left(\boldsymbol{\Phi}-\hat{\boldsymbol{\Phi}}_B\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Phi}-\hat{\boldsymbol{\Phi}}_B\right)\right]\right\},$$

em que:

$$D = \beta + \frac{\left(\mathbf{Y}'\mathbf{Y} + \boldsymbol{\mu}'\mathbf{I}\boldsymbol{\mu}\right) - \left(\mathbf{X}'\mathbf{Y} + \mathbf{I}\boldsymbol{\mu}\right)' \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\right)^{-1} \left(\mathbf{X}'\mathbf{Y} + \mathbf{I}\boldsymbol{\mu}\right)}{2},$$

$$\hat{\boldsymbol{\Phi}}_B = \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\right)^{-1}\left(\mathbf{X}'\mathbf{Y} + \mathbf{I}\boldsymbol{\mu}\right) \text{ e } \boldsymbol{\Sigma} = \mathbf{X}'\mathbf{X} + \mathbf{I}.$$

Para fazer inferências sobre os parâmetros de interesse é necessário obter suas distribuições marginais a *posteriori.* Estas distribuições são obtidas por meio da integração da distribuição a *posteriori* em relação a todos os parâmetros, exceto o de interesse.

Na maioria dos casos, essas integrais são complexas e não apresentam soluções exatas. Para contornar este problema, recomenda-se a utilização de algoritmos MCMC (Markov Chain Monte Carlo), como o amostrador de Gibbs (GELFAND; SMITH 1990) e/ou Metropolis-Hastings (HASTINGS, 1973).

2.5 Previsões k-passos a frente via MCMC

Um dos objetivos mais importantes da análise de séries temporais é fazer previsão de valores futuros da série (WEI, 1994). 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 resolução de uma integral múltipla em relação a todos os parâmetros da distribuição conjunta das observações futuras e dos parâmetros, condicionada aos dados passados (BARRETO; ANDRADE, 2004).

Considerando o modelo estatístico AR(p) para dados em painel de um valor futuro, $Y_{i(t+1)} = \mu_i + \phi_{i1} Y_{i(t)} + \phi_{i2} Y_{i(t-1)} + ... + \phi_{ip} Y_{i(t+1-p)} + e_{i(t+1)}$, em que $e_{i(t+1)} \sim N(0, \sigma_e^2)$, a função de verossimilhança referente a todos os indivíduos i (i=1,2,...,m), supostos independentes, sob a forma matricial é dada por:

$$L\left(\mathbf{Y}_{(t+1)} \mid \Phi, \sigma_e^2, \mathbf{Y}\right) \propto \sigma_e^{2\,\frac{-m}{2}} \exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{Y}_{(t+1)} - \mathbf{X}\Phi\right)'\left(\mathbf{Y}_{(t+1)} - \mathbf{X}\Phi\right)\right\},$$

em que: $\mathbf{Y}_{(t+1)} = \left[y_{1(t+1)}, y_{2(t+2)}, \ldots, y_{m(t+1)}\right]'_{mx1}$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \mathbf{X}_m \end{bmatrix}_{m \times mp} \quad \text{e} \quad \mathbf{X}_i = [1, y_{i(t)}, y_{i(t-1)}, ..., y_{i(t+1-p)}]_{1 \times p}.$$

Desta forma, a distribuição preditiva é dada pela seguinte integral:

$$P\left(\mathbf{Y}_{(t+1)} \mid \mathbf{Y}\right) \propto \int_{\sigma_e^2} \int_{\Phi} \sigma_e^{2\,\frac{-m}{2}} \exp\left\{ -\frac{1}{2\sigma_e^2} \left(\mathbf{Y}_{(t+1)} - \mathbf{X}\boldsymbol{\Phi}\right)' \left(\mathbf{Y}_{(t+1)} - \mathbf{X}\boldsymbol{\Phi}\right) \right\} P\left(\boldsymbol{\Phi}, \sigma_e^2 \mid \mathbf{Y}\right) d\boldsymbol{\Phi} d\sigma_e^2$$

A obtenção do valor de $P\left(\mathbf{Y}_{(t+1)} \mid \mathbf{Y}\right)$ por meio analítico é quase sempre inviável. Assim, para a obtenção de uma aproximação para a integral acima faz-se necessário o uso de métodos MCMC. Segundo Heckman e Leamer (2001) mediante a utilização da técnica MCMC, tem-se:

$$\mathbf{Y}_{(n+1)}^{(q)} \mid \mathbf{Y} \sim N\left(\mathbf{X}\boldsymbol{\Phi}^{(q)}, \sigma_e^{2(q)}\mathbf{I}\right),$$

em que: **I** é uma matriz identidade de dimensões m(p+1) x m(p+1).

De forma geral, se constatada a convergência dos algoritmos MCMC, podemos assumir que o conjunto de valores gerados para esta distribuição Normal, provenientes da q-ésima iteração dos algoritmos Metropolis-Hastings e/ou Gibbs *sampler*, constituem a distribuição preditiva para um dado futuro, cuja estimativa do valor predito é $\hat{\mathbf{Y}}_{(t+1)}$. 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 valores futuros da série, porém para esta implementação é necessário obedecer a um processo iterativo, fundamentado na ordem de geração dos valores, ou seja, para gerar a distribuição de $\mathbf{Y}_{(t+2)}$, deve-se anteriormente gerar a distribuição de $\mathbf{Y}_{(t+1)}$, e assim sucessivamente até a predição $\mathbf{Y}_{(t+k)}$.

2.6 Critérios para a seleção de modelos

Seja na análise de séries temporais ou em qualquer análise de dados, diversos modelos podem ser utilizados para representar o comportamento de um mesmo conjunto de dados (WEI, 1994). Entretanto, nem sempre a escolha do melhor modelo é fácil. Para auxiliar nesta escolha, diversos critérios para comparação de modelos são descritos na literatura. Dentre estes, os que possuem maior destaque são os critérios de informação de Akaike (1973, 1974, 1978, 1979) e o de Schwartz (1978) também conhecido como critério de informação bayesiano (*Bayesian Information Criterion - BIC*).

2.6.1 Critério de informação de Akaike (*Akaike's Information Criterion – AIC*)

Considere que um modelo estatístico com M parâmetros é ajustado aos dados. Para avaliar a qualidade do ajuste do modelo, Akaike (1973, 1974) definiu a seguinte medida:

$$\mathrm{AIC(M)} = -2\ln(\mathrm{L}(\mathrm{y} \mid \hat{\theta})) + 2\mathrm{M}\,,$$

em que: M é o número de parâmetros do modelo ajustado.

Posteriormente, Akaike (1978, 1979) desenvolveu outra medida, baseada em inferência bayesiana, definida como:

$$\mathrm{BAIC(M)} = \mathrm{n}\ln(\mathrm{L}(\mathrm{y} \mid \hat{\theta})) - (\mathrm{n} - \mathrm{M})\ln\left(1 - \frac{\mathrm{M}}{\mathrm{n}}\right) + \mathrm{M}\ln\mathrm{n} + \mathrm{M}\ln\left[\left(\frac{\mathrm{L}(\mathrm{y} \mid \hat{\theta})}{\hat{\sigma}_{\mathrm{Y}}^2}\right)/\mathrm{M}\right].,$$

em que: n: número de observações disponíveis para o ajuste; $\hat{\sigma}_{\mathrm{Y}}^2$: é a variância amostral da série; e M o número de parâmetros do modelo. Segundo Akaike (1978) o critério *BAIC* tem melhores propriedades.

2.6.2 Critério de informação de Schwartz (*Bayesian Information Criterion - BIC*)

Utilizando uma ideia similar ao critério bayesiano de Akaike, Schwartz (1978), sugeriu outra medida para a seleção de modelos dada pela seguinte expressão:

$$BIC(M) = n\ln(L(y \mid \hat{\theta})) + M\ln(n),$$

em que: n : número de observações disponíveis para o ajuste; e M o número de parâmetros do modelo.

Em ambos as medidas apresentadas o modelo que apresentar menor valor é considerado o mais adequado para representar os dados em estudo. Para informações detalhadas, bem como outros critérios de seleção de modelos pode-se consultar (HANNAN; QUINN, 1979; STONE, 1979).

2.7 Análise de agrupamento (*cluster analysis*)

A análise de agrupamento tem por objetivo criar subgrupos distintos e homogêneos a partir de um dado conjunto de indivíduos (FRANZÉN, 2008).

Geralmente, na maior parte dos estudos práticos, a análise de agrupamentos é realizada por meio de métodos determinísticos os quais se utilizam de medidas entre os indivíduos como base para a formação de grupos homogêneos. Estes métodos são divididos, de maneira geral, em métodos hierárquicos e não hierárquicos (FERREIRA, 2008). Dentre os não hierárquicos, os que apresentam grande destaque são os métodos de Tocher e o conhecido como k-médias. Esses métodos às vezes são chamados de métodos de otimização, uma vez que fazem uso de uma função objetivo como critério de agrupamento. Neste trabalho, com o intuito de obter conjuntos de genes

homogêneos, serão abordados apenas os principais métodos hierárquicos e o método de otimização de Tocher.

2.7.1 Métodos de agrupamento hierárquicos

Nos métodos hierárquicos os indivíduos em estudo são classificados em grupos em diferentes etapas, de modo hierárquico, produzindo uma árvore de classificação, os dendrogramas (BUSSAB; MIAZAKI; ANDRADE, 1990).

Os métodos hierárquicos são também divididos em aglomerativos e divisivos. Dentre os aglomerativos, citam-se o do vizinho mais próximo (*Single Linkage Method*); o do vizinho mais distante (*Complete Linkage Method*); o da ligação média (*Average Linkage*), ponderado ou não; e o proposto por Ward (1963). Dentre os divisivos, o mais conhecido é o de Edwards e Cavalli-Sforza (1965).

2.7.1.1 Método do vizinho mais próximo

Segundo Sneath e Sokal (1973) o método do vizinho mais próximo foi introduzido na Taxonomia. Neste método, é estabelecido um dendrograma pelos indivíduos com maior similaridade, sendo a distância entre um indivíduo k e um grupo, formado pelos indivíduos i e j, dada por:

$$d_{(ij)k} = \min\{d_{ik}, d_{jk}\}$$

A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = \min\{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$$

ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k e l) é dada pela menor distância entre os pares de indivíduos (i e k), (i e l), (j e k) e (j e l).

2.7.1.2 Método do vizinho mais distante

Neste método o dendrograma é estabelecido pelos indivíduos com menor similaridade, sendo a distância entre um indivíduo k e um grupo, formado pelos indivíduos i e j, dada por:

$$d_{(ij)k} = \max\{d_{ik}, d_{jk}\}$$

A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = \max\{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$$

ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k e l) é dada pela maior distância entre os pares de indivíduos (i e k), (i e l), (j e k) e (j e l).

2.7.1.3 Método da ligação média entre grupos ou UPGMA (Unweighted pair-group method using arithmetic averages)

Como regra geral, a construção do dendrograma é estabelecida pelo indivíduo de maior similaridade. Entretanto, a distância entre um indivíduo k e um grupo formado pelos indivíduos i e j, é dada por:

$$d_{(ij)k} = \frac{d_{ik} + d_{jk}}{2},$$

ou seja, $d_{(ij)k}$ é dada pela média do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k).

A distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k, l e m) é dada por:

$$d_{(ij)(klm)} = \frac{d_{ik} + d_{il} + d_{im} + d_{jk} + d_{jl} + d_{jm}}{6}$$

ou seja, é determinada pela média entre os elementos do conjunto, cujos elementos são distâncias entre pares de indivíduos de grupos (i e k),(i e l), (i e m), (j e k), (j e l) e (j e m).

2.7.1.4 Método de Ward

De acordo com o método de Ward (1963), também conhecido como método da "Mínima Variância", os grupos são formados por meio da maximização da homogeneidade dentro dos grupos, isto é, unem-se dois grupos R e S que minimizam o incremento na soma do quadrado do erro (SSE). Assim, como descrito em Ferreira (2008), se em um determinado estágio do processo de agrupamento, tivermos $k \leq n$ grupos e se o i-ésimo objeto do l-ésimo grupo for representado por $\mathbf{y}_i^{(l)}$, l=1,2, ..., k *e* i=1,2, ..., n_l, pode-se definir a soma de quadrados do erro no l-ésimo grupo por:

$$SSE_l = \sum_{i=1}^{n_l} \left(\mathbf{y}_i^{(l)} - \overline{\mathbf{y}}_\cdot^{(l)}\right)' \left(\mathbf{y}_i^{(l)} - \overline{\mathbf{y}}_\cdot^{(l)}\right),$$

em que n_l é o número de objetos no l-ésimo grupo e $n = \sum_{l=1}^{k} n_l$.

Desta forma, quando consideram-se dois grupos quaisquer, denotados por R e S, deve-se aglomerar o par que minimiza a troca na SSE, provocada pela junção dos grupos R e S em um único grupo, denotada por ΔSSE_{RS}. Assim, a troca em SSE provocada pela junção dos grupos é dada por:

$$\Delta SSE_{RS} = SSE_{RS} - SSE_R - SSE_S,$$

em que:

$$SSE_{RS} = \sum_{i=1}^{n_{RS}} \left(\mathbf{y}_i^{(RS)} - \overline{\mathbf{y}}_\cdot^{(RS)}\right)' \left(\mathbf{y}_i^{(RS)} - \overline{\mathbf{y}}_\cdot^{(RS)}\right),$$

em que: $n_{RS} = n_R + n_S$ e $\bar{\mathbf{y}}_{.}^{(RS)} = \frac{\left(n_R \bar{\mathbf{y}}_{.}^{(R)} + n_S \bar{\mathbf{y}}_{.}^{(S)}\right)}{n_R + n_S}$, são o tamanho e o centroide do novo grupo RS, respectivamente.

Ward (1963) demonstrou que a troca na soma de quadrados dos erros, provocada pela junção dos grupos R e S, depende diretamente da distância quadrática entre os centroides dos grupos correspondentes, que pode ser simplificada em

$$\Delta SSE_{RS} = \frac{n_R n_S}{n_R + n_S}\left(\bar{\mathbf{y}}_{i}^{(R)} - \bar{\mathbf{y}}_{.}^{(S)}\right)'\left(\bar{\mathbf{y}}_{i}^{(R)} - \bar{\mathbf{y}}_{.}^{(S)}\right).$$

2.7.2 Métodos de otimização

Nos métodos de otimização realiza-se a partição do conjunto de indivíduos em subgrupos não vazios e mutuamente exclusivos por meio da maximização ou minimização de alguma medida preestabelecida. Dois métodos de otimização são comumente utilizados em genética e melhoramento, o proposto por Tocher, citado por Rao (1952) e o de Tocher modificado (VASCONCELOS et al., 2007).

2.7.2.1 Métodos de Tocher original e modificado

Para a utilização do método de Tocher faz-se necessário a utilização da matriz de dissimilaridade, ou seja, uma matriz na qual a distância entre os elementos é quantificada, sobre a qual é identificado o par de indivíduos mais similares. Esses indivíduos formarão o grupo inicial. A partir deste ponto a inclusão de novos indivíduos é avaliada adotando-se o critério de que a distância média intragrupo deve ser menor que a distância média intergrupo (CRUZ;

CARNEIRO, 2003). A medida de dissimilaridade utilizada depende da natureza da variável em estudo.

A entrada de um indivíduo em um grupo sempre aumenta o valor médio da distância dentro do grupo. Assim, pode-se tomar a decisão de incluir o indivíduo em um grupo por meio da comparação entre o acréscimo no valor médio da distância dentro do grupo e um nível máximo permitido, que pode ser estabelecido arbitrariamente, ou adotado, o valor máximo (θ) da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo. Assim, a inclusão, ou não, do indivíduo k no grupo é, então, feita considerando:

- Se $\frac{d_{(grupo)k}}{n} \leq \theta$, inclui-se o indivíduo k no grupo;
- Se $\frac{d_{(grupo)k}}{n} > \theta$, o indivíduo k não é incluído no grupo.

em que: n representa o número de indivíduos que constitui o grupo original.

Neste caso, a distância entre o indivíduo k e o grupo formado pelos indivíduos ij é dada por $d_{(ij)k} = d_{ik} + d_{jk}$ (CRUZ; CARNEIRO, 2003).

No método de Tocher modificado o valor de (θ) é obtido de maneira sequencial, isto é, a cada início do processo de formação de grupo o valor de (θ) é obtido novamente da matriz de dissimilaridade composta apenas pelos indivíduos não agrupados. Para mais detalhes sobre a modificação (VASCONCELOS et al., 2007).

2.8 Determinação do número de grupos para métodos hierárquicos

A análise de agrupamento constitui uma das mais importantes técnicas multivariada quando se objetiva a classificação de diversos indivíduos em um número reduzido de grupos. Entretanto, a técnica baseada em métodos

hierárquicos não fornece um critério objetivo para determinar o número "ótimo" de grupos em que os indivíduos em estudo devam ser alocados. Algumas propostas com o intuito de preencher esta lacuna são descritas na literatura. Mojema (1977) propôs um critério baseado na maior amplitude das distâncias de junção dos grupos formados com o objetivo de determinar um número k que otimize a qualidade do agrupamento dos dados. Deve-se escolher o número de grupos dado pelo primeiro estágio do dendograma no qual:

$$\alpha_j > \overline{\alpha} + \psi S_\alpha ,$$

em que: j=1,2,...,n; α_j é o valor da distância para o estágio de junção correspondente a n-j+1 grupos; $\overline{\alpha}$ e S_α são a média e o desvio padrão dos $\alpha's$ e; ψ é uma constante que de acordo com Milligan e Cooper (1995) deve assumir valor 1,25.

Sharma (1996) apresentou a estatística *RMSSTD (Root mean square standard deviation)* a qual permite obter o número ótimo de *clusters*. Segundo Silveira (2010), a estatística *RMSSTD* tem sido utilizada com maior frequência em trabalhos na área de Biometria como nos trabalhos de Araújo (2008) e Cecon et al. (2008).

A estatística *RMSSTD* é utilizada para calcular a homogeneidade dos agrupamentos. Assim, quanto maior for o número de grupos formados menores serão os valores de *RMSSTD*. A estratégia de determinação do número ótimo de *clusters* consiste em encontrar o ponto de máxima curvatura da trajetória, não linear, do *RMSSTD* em função do número de grupos. O ponto de máxima curvatura indica um limiar entre uma fase de decréscimo e uma fase de estabilização, assim, o aumento do número de grupos após este ponto não altera significativamente os valores de *RMSSTD*.

O comportamento do *RMSSTD* em relação ao número de grupos (NG) pode ser descrito por um modelo exponencial, $RMSSTD = a(\mathrm{NG})^{-b} + e$, em que a

e b são os parâmetros deste modelo, e NG corresponde ao número de grupos formados. Assim, como no trabalho de Cecon et al. (2008) e Silveira (2010), o número "ótimo" de grupos (X_0), pode ser determinado geometricamente por meio da interseção desta curva com uma reta, de forma que a maior distância entre elas corresponda ao ponto em questão. O cálculo de *RMSSTD*, para cada novo grupo formado é realizado por meio da expressão:

$$RMSSTD_k = \sqrt{\frac{SQ_1 + SQ_2 + \ldots + SQ_p}{gl_1 + gl_2 + \ldots + gl_p}}$$

em que: $SQ_j = \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2$ é a soma de quadrado da j-ésima variável calculada considerando as n observações presentes em cada novo *cluster* k e gl_j é o número de graus de liberdade da variável j em questão.

2.9 Dados de MTS (Microarray Time Series)

Bioinformática é a área que estuda o projeto e implementação de novos métodos computacionais para auxiliar pesquisadores de áreas como Biologia Molecular, Bioquímica, entre outras, principalmente na análise de grandes quantidades de dados biológicos. Os avanços tecnológicos e da pesquisa em áreas como genômica, transcriptoma e proteômica têm produzido enormes massas 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 et al., 2007a).

Alguns dos problemas mais importantes estudados nesta área estão ligados à compreensão do funcionamento do sistema celular, interações

genéticas e desenvolvimento de doenças genéticas no âmbito molecular. Desta forma, o monitoramento da mudança no padrão de expressão gênica através do tempo, proporciona a possibilidade de desvendar os padrões de respostas celulares (ANDROULAKIS; YANG; ALMON, 2007).

A modelagem em questão é feita a partir de dados provenientes de técnicas de expressão, dentre as quais se destaca o *microarray* que surgiu em meados dos anos 90 (DERISI et al., 1996; SCHENA et al., 1995; SHALON; SMITH; BROWN, 1996). O *microarray* é uma técnica, que tem-se mostrado bastante útil para a quantificação simultânea dos níveis de expressão gênica, inclusive ao longo do tempo, de milhares de genes.

De forma sucinta, a técnica de *microarray*, consiste na alocação de sequências de cDNA conhecidas ou pequenos oligonucleotídeos em posições especificas de uma lâmina de vidro ou em uma membrana de *nylon*, que são hibridizados contra cDNA marcados. No caso da utilização de membranas de *nylon*, os cDNA são marcados radioativamente e utiliza-se apenas um tipo biológico por membrana. Já quando se utiliza lâminas de vidro, utilizam-se dois tipos de amostras biológicas que são marcadas por fluorescência (*dyes cy3 e cy5*). Após essa fase, a membrana ou lâmina deve passar por um processo de digitalização. No caso das lâminas de vidro, os fluorocromos são excitados emitindo sinais luminosos captados por um *scanner*. Os dados originais de *microarray* são imagens que representam os níveis de expressão dos genes fixados no material utilizado. Estas imagens devem ser analisadas por um *software* específico que gere uma tabela de dados numéricos contendo os valores de intensidade (ESTEVES, 2007).

A variável resposta em um experimento de *microarray* é razão entre a intensidade de luz emitida pelos genes do grupo tratado e do grupo controle. Para facilidade de interpretação, toma-se o $\log_2$ desta razão, o qual é denominado *fold-change*. Estes valores são calculados para cada gene, sendo

que um valor de *fold-change* positivo indica que o gene se expressa mais no grupo tratado (gene *up-regulated* para tratamento), um valor igual a zero indica que o gene é igualmente expresso em ambos os grupos, e um valor negativo indica que o gene se expressa mais no grupo controle (gene *down-regulated* para tratamento) (MORAIS et al., 2010).

A Figura 1 representa esquematicamente o procedimento experimental para lâminas de vidro usando moléculas de cDNA para fixação.

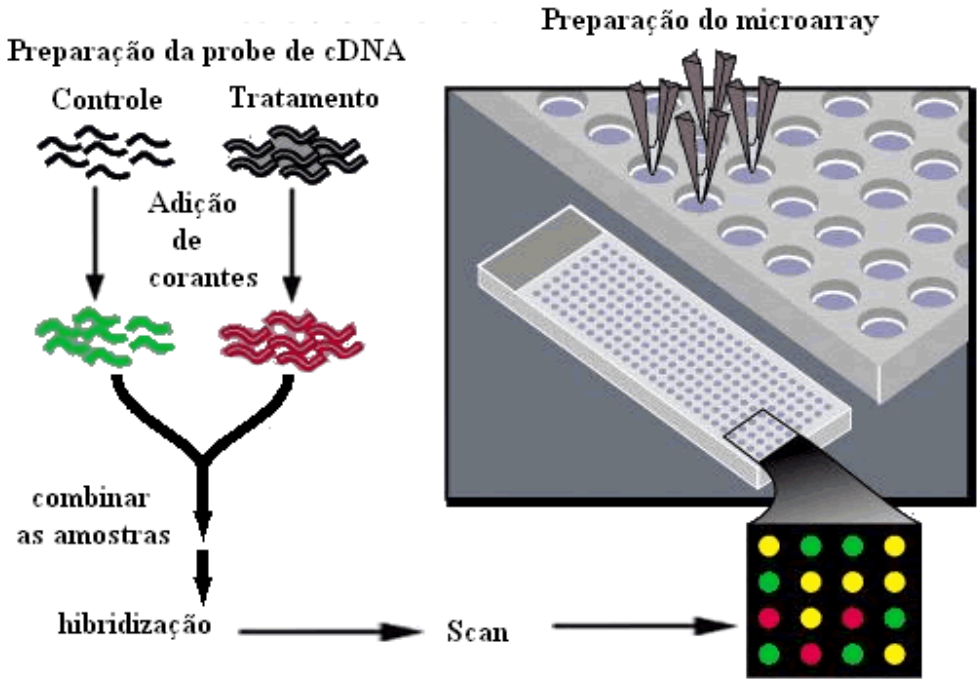

Figura 1 Esquema de um experimento de *microarray*
Fonte: Brainarray…(2011).

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 órgãos e tecidos, drogas, instantes de tempo. Quando está se tratando

de diferentes instantes de tempo, a modelagem pode ser especialmente útil para a identificação de conjuntos de genes com comportamentos semelhantes em determinados instantes temporais. Assim, quando estes dados temporais são avaliados, os mesmos podem ser denominados de *Microarray Time Series* (*MTS)* (MUKHOPADHYAY; CHATTERJEE, 2007). 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.

Alguns trabalhos, descritos na literatura, utilizaram modelos de séries temporais para o ajuste dos dados provenientes de experimentos de *Microarray Time Series*, dentre estes, pode-se citar o trabalho de Fujita et al. (2007a) em que o autor comparou as metodologias *SVR* (*Support Vector Regression*), *DVAR* (*Dynamic Vector Autoregressive Model*) e *SVAR* (*Sparse Vector Autoregressive Model*), fundamentadas na modelagem de vetores autorregressivos (*VAR*), 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). Além desse, Morais et al. (2010) utilizaram a análise bayesiana do modelo autorregressivo de primeira ordem para dados em painel para descrever o comportamento de alguns genes, cujas trajetórias de expressão foram consideradas homogêneas, utilizados no mesmo estudo de Fujita et al. (2007a).

3 CONCLUSÃO

Apresentou-se. com o objetivo de explanar todas as técnicas utilizadas no desenvolvimento dos artigos, toda a teoria dos modelos autorregressivos para dados em painel, bem como uma breve introdução sobre inferência bayesiana e os principais métodos de simulação *MCMC*. Além disso, apresentou-se também a abordagem bayesiana do modelo autorregressivo para dados em painel. Posteriormente, foram abordados alguns critérios para a seleção de modelos, métodos de agrupamento hierárquicos e de otimização e a determinação do número “ótimo” de grupos para os métodos hierárquicos. Finalmente, realizou-se uma breve introdução de como são obtidos os dados de *MTS* (*Microarray Time Series*). Acredita-se que a apresentação destas técnicas é de fundamental importância para o entendimento das análises realizadas.

REFERÊNCIAS

AKAIKE, H. A bayesian analysis of the minimum **AIC procedure. Annals of the Institute of Statistical Mathematics**, Tokio, v. 30, p. 9-14, 1978.

AKAIKE, H. A bayesian extension of the minimum AIC procedure of autoregressive model fitting. **Biometrika**, London, v. 66, n. 2, p. 237-242, 1979.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, Boston, v. 19, n. 6, p. 716-723, Dec. 1974.

AKAIKE, H. Information theory and a extension of the maximum likelihood principle. In: INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY, 2., 1973, Budapest. **Procedings…** Budapest: Akademiai Kiado, 1973. p. 367-381.

ANDROULAKIS, I. P.; YANG, E.; ALMON, R. R. Analysis of time series gene expression data: methods, challenges and opportunities. **The Annual Review of Biomedical Engennering**, Palo Alto, v. 9, p. 225-228, 2007.

ARAÚJO, M. N. M. **Análise de sobrevivência do tomateiro a *Phytophthora infestans***. 2008. 53 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, MG, 2008.

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

BALGOBIN, N.; PETRUCELLI, J. D. A Baysian analysis of autoregressive time series panel data. **Journal of Business and Economic Statistics**, Boston, v. 1, n. 1,
p. 328-334, 1997.

BAR-JOSEPH, Z. Analyzing time series gene expression data. **Bioinformatics**, Oxford, v. 20, p. 2493-2503, 2004.

BAR-JOSEPH, Z. et al. Continuous representations of time series gene expression data. **Journal of Computational Biology**, New York, v. 3, p. 341–356, 2003.

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

BRAINARRAY molecular and behavioral neuroscience institute. Disponível em: http://brainarray.mbni.med.umich.edu/Brainarray/Resources/Method.asp. Acesso em: 21 dezembro 2010.

BUSSAB, W. O.; MIAZAKI, É. S.; ANDRADE, D. F. **Introdução à análise de agrupamentos**. São Paulo: ABE, 1990. 105 p.

CAMERON, A. C.; TRIVEDI, P. K. **Microeconometrics**: methods and aplications. Cambridge: Cambridge University, 2005. 1034 p.

CASELLA, G.; GEORGE, E. Explaining the gibbs sampler. **The American Statistician**, Alexandria, v. 46, p. 167-157, 1992.

CECON, P. R. et al. Análise de medidas repetidas na avaliação de clones de café 'Conilon'. **Pesquisa Agropecuária Brasileira**, Brasília, v. 43, p. 1171-1176, 2008.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **The American Statistician**, Alexandria, v. 49, p. 327-335, 1995.

COSTA, I. G.; CARVALHO, F. A. T.; SOUTO, M. C. P. Comparative analysis of clustering methods for gene expression time course Data. **Genetics and Molecular Biology**, Ribeirão Preto, v. 27, p. 623-631, 2004.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. 2. ed. Viçosa, MG: UFV, 2003. 585 p.

DERISE, J. et al. Use of a cDNA *microarray* to analyse gene expression patterns in human cancer. **Nature Genetics**, New York, v. 14, p. 457-460, 1996.

EDWARDS, A. W. F.; CAVALLI-SFORZA, L. L. A method for cluster analysis**. Biometrics**, Washington, v. 21, p. 362-375, 1965.

EISEN, M. B. et al. Cluster analysis and display of genome-wide expression patterns. **Proceedings of the National Academy of Sciences of America,** Washington, v. 95, p. 14863-14868, 1998.

ERNST, J.; NAU, G. J.; BAR-JOSEPH, Z. Clustering short time series gene expression data. **Bioinformatics**, Oxford, v. 21, p. 159-168, 2005.

ESTEVES, G. H. **Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado**. 2007. 174 p. Tese (Doutorado em Bioinfomática) – Universidade de São Paulo, São Paulo, 2007.

FERREIRA, D. F. **Estatística multivariada**. Lavras: UFLA, 2008. 662 p.

FRANZÉN, J. **Bayesian cluster analysis**: some extensions to non-standard situations. 2008. 161 f. Thesis (Doctoral in Statistics) – University Stockholm, Stockholm, 2008.

FUJITA, A. et al. GEDI: a user-friendly toolbox for analysis of large-scale gene expression data. **BMC Bioinformatics**, London, v. 8, p. 457, 2007a.

FUJITA, A. et al. Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. **Bioinformatics**, Oxford,
v. 23, p. 1623–1630, 2007b.

GELFAND, A. E.; SMITH, A. F. M. Sampling based approaches to calculating marginal densities**. Journal of the American Statistical Association**, Alexandria**,**
v. 85, p. 398-409, 1990.

GEWEKE, J. **Evaluating the accurary of sampling-based approaches to the calculation of posterior moments.** New York: Oxford University, 1992. p. 625-631**. (**Bayesian Statistics, 4).

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. **Journal of the Royal Statistical Society: Serie B,** Oxford, v. 41**,** p. 190-195, 1979.

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

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

HEIDELBERGER, P.; WELCH, P. Simulation run lenght control in the presence of an initial transient. **Operations Research**, Landing, v. 31, p. 1109-1144, 1993.

HOLTZ-EAKIN, D.; NEWEY, W.; ROSEN, H. S. Estimating vector autoregressions with panel data. **Econometrica**, New York, v. 56, p. 1371-1395, 1988.

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

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

METROPOLIS, N. et al. Equation of state calculations by fast computing machine. **Journal of Chemical Physics**, New York, v. 21, p. 1087-1091, 1953.

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

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, Colorado Springs, v. 50, p. 159-179, 1985.

MOJEMA, R. Hierarquical grouping methods and stopping rules: na evaluation. **Computer Journal**, Trier, v. 20, p. 359-363, 1977.

MORAIS, T. S. S. et al. Análise bayesiana de sensibilidade do modelo AR(1) para dados em painel: uma aplicação em dados temporais de *microarrays*. **Revista Brasileira de Biometria**, São Paulo, v. 28, p. 171-192, 2010.

MORETTIN, P. A.; TOLOI, C. M. C. **Analise de series temporais**. 2. ed. São Paulo: E. Blucher, 2006. 535 p.

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

NANDRAM, B.; PETRUCCELLI, J. D. **Journal of Business & Economic Statistics**, Boston, v. 15, p. 328-34, 1997.

PAEZ, M.; GAMERMAN, D. **Modelagem de processos espaço temporais**. 11. ed.· São Paulo: ABE, 2005. 102 p.

PAULINO, C. D.; TURKMAN, M. A.; MURTEIRA, B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

RAFTERY, A. E.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J. M. et al. **Bayesian statistics 4.** Oxford: Oxford University, 1992. p. 763-773.

RAMONI, M. F.; SEBASTIANI, P.; KOHANE, I.S. Cluster analysis of gene expression dynamics. **Proceedings of the National Academy of Sciences of America,** Washington, v. 99, p. 9121-9126, 2002.

RAO, R. C. **Advanced statistical methods in biometric research**. New York: J. Wiley, 1952. 39 p.

R Development Core Team. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2008. Disponível em: <http://www.R- project.org>. Acesso em: 22 fev. 2011.

RIVERO, S. et al. Pecuária e desmatamento: uma análise das principais causas direta do desmatamento na Amazônia. **Nova Economia**, Belo Horizonte, v. 9, p. 41-66, 2009.

SANTOS, M. J. Dinâmica temporal da criminalidade: mais evidências sobre o efeito inércia nas taxas de crimes letais nos estados brasileiros. **Economia**, Brasília, v. 10, p. 169-194, 2009.

SCHENA, M. et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science**, Washington, v. 270, p. 467-470, 1995.

SCHLIEP, A.; SCHONHUTH, A.; STEINHOFF, C. Using hidden markov models to analyze gene expression time course data. **Bioinformatics**, Oxford, v. 19, p. 264-272, 2003.

SCHWARTZ, G. Estimating the dimension of a model. The **Annals of Statistics.** Philadelphia, v. 6, p. 461-464, 1978.

SHALON, D.; SMITH, S. J.; BROWN, P. O. A DNA microaaray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. **Genome Research**, New York, v. 6, p. 639-645, 1996.

SHARMA, S. **Applied multivariate techniques**. NewYork: J. Wiley, 1996. 493 p.

SILVA, F. F. et al. Bayesian analysis of autoregressive panel data model: application in genetic evaluation of beef cattle. **Scientia Agrícola**, Piracicaba, v. 68, p. 237-245, 2011.

SILVA, F. F. et al. Comparação bayesiana de modelos de previsão de diferenças esperadas nas progênies no melhoramento genético de gado Nelore. **Pesquisa Agropecuária Brasileira**, Brasília, v. 43, p. 37-45, 2008a.

SILVA, F. F. et al. Previsão bayesiana de valores genéticos de touros por meio do modelo autorregressivo para dados em painel. **Arquivo Brasileiro de Medicina Veterinária *e* Zootecnia**, Belo Horizonte, v. 60, p.1166-1173, 2008b.

SILVEIRA, F. G. **Classificação multivariada de modelos de crescimento para grupos genéticos de ovinos de corte.** 2010. 61 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, MG, 2010.

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

SNEATH, P. H.; SOKAL, R. R. **Numerical taxonomy**: the principles and practice of numerical classification. San Francisco: W. H. Freeman, 1973. 573 p.

SOUZA, T. R. V.; LEITE FILHO, P. A. M. Análise por dados em painel do status de saúde no Nordeste Brasileiro. **Revista de Saúde Pública**, São Paulo, v. 42, p. 96-804, 2008.

STONE, M. Comments on model selection criteria of Akaike and Schwartz. **Journal of the Royal Statistical Society: Serie B**, Oxford, v. 41, p. 276-278, 1979.

UNIVERSIDADE FEDERAL DE LAVRAS. Biblioteca. **Manual de normalização e estrutura de trabalhos acadêmicos**: TCC, monografias, dissertações e teses. Lavras, 2010. Disponível em: <http://www.biblioteca.ufla.br/site/index.php>. Acesso em: 23 mar. 2011.

VASCONCELOS, E. S. et al. Método alternativo para análise de agrupamento. **Pesquisa Agropecuária Brasileira**, Brasília, v. 42, p. 1421-1428, 2007.

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v. 58, p. 236-244, 1963.

WEI, W. W. S. **Time series analysis**: univariate and multivariate methods. New York: A. Wesley, 1994. 478 p.

YAMAGUCHI, R. et al. Finding module-based gene networks in time-course gene expression data with state space models. **IEEE Signal processing magazine**, New York, v. 24, p. 37-46, 2007.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 Análise de agrupamento para dados de expressão gênica temporal: uma aplicação em dados em painel

RESUMO

Este trabalho teve por objetivo avaliar qual a melhor alternativa, dentre os métodos de agrupamentos hierárquicos (Ward) e de otimização (Tocher), na formação de grupos homogêneos de séries de expressão gênica para posterior ajuste de modelos autorregressivos, AR(p), para dados em painel. Para tanto, utilizaram-se as estimativas dos parâmetros de modelos autorregressivos previamente ajustados as séries individuais (de cada gene) de dados *MTS (Microarray Time Series*) como variáveis no processo de agrupamento. Além disso, também ajustou-se o modelo AR(p) para dados em painel a fim de realizar previsões da expressão gênica dentro de cada grupo formado. Os dados utilizados referem-se à expressão de genes que atuam sobre ciclo celular de *Saccharomyces cerevisiae*. Tais dados correspondem a 114 genes, sendo que, cada um deles apresentava 10 valores de *fold-change* (medida da expressão) ao longo do tempo (0, 15, 30, ..., 135 min).

Palavras-chave: Série temporal, modelo autorregressivo, *microarray*.

Cluster analysis for temporal gene expression data: an application to panel data

ABSTRACT

The objective of this study was to determine which one of the hierarchical clustering (Ward) and optimization methods (Tocher) is the best to form homogeneous groups of gene expression sets for subsequent autoregressive model fitting, AR (p) for panel data. For this purpose, estimates of the autoregressive model parameters previously adjusted to individual series (of each gene) of MTS data (Microarray Time Series) were used as variables in the clustering process. In addition, the model AR (p) was also fitted to panel data to make predictions of the gene expression within each group formed. Data of the expression of genes that control the cell cycle of *Saccharomyces cerevisiae* were used, corresponding to 114 genes, of which each had 10 fold-change values (measure of expression) over time (0, 15, 30, ..., 135 min.)

Index terms: time series, autoregressive model, microarray.

1 INTRODUÇÃO

A análise de dados de expressão gênica identificada ao longo do tempo, os quais são denominados *Microarray Time Series* (*MTS*), tem possibilitado o entendimento de diversos processos biológicos (MUKHOPADHYAY; CHATTERJEE, 2007), pois o conhecimento de grupos de genes que se expressam de forma similar possibilita inferir a respeito de funções e mecanismos reguladores dos mesmos (COSTA; CARVALHO; SOUTO, 2004).

Apesar dos métodos de agrupamento hierárquicos e os de otimização (EISEN et al., 1998) serem extremamente utilizados em problemas biológicos, os mesmos não levam em consideração a natureza sequencial das observações. Para contornar essa situação, foram desenvolvidos alguns métodos baseados no ajuste de modelos específicos de regressão. Dentre estes, destacam-se os de agrupamentos que têm como base a dinâmica do padrão de expressão gênica (RAMONI; SEBASTIANI; KOHANE (2002), os modelos de Markov oculto (SCHLIEP; SCHONHUTH; STEINHOFF, 2003) e o agrupamento por meio de representações contínuas do tipo B-splines (BAR-JOSEPH et al., 2003). Porém, apesar de úteis, segundo Bar-Joseph (2004) os mesmos não são adequados para experimentos relativamente pequenos, ou seja, com menos de 10 observações temporais por gene.

Portanto, uma forma prática de associar estes métodos usuais de agrupamentos (como o Ward e Tocher) às análises de dados *MTS*, é considerar como variáveis na aplicação de tais métodos as estimativas de parâmetros de modelos que consideram esta natureza sequencial das observações, como por exemplo, os modelos autorregressivos AR(p). Por meio desta metodologia, é possível obter a formação de grupos de genes homogêneos em relação as suas expressões temporais.

Após a obtenção destes grupos, é possível ainda ajustar modelos AR(p) para dados em painel separadamente para cada grupo, possibilitando assim o aumento da precisão das estimativas dos parâmetros em relação às análises individuais de cada série (LIU; TIAO, 1980; MORAIS et al., 2010; SILVA et al., 2011), e consequentemente o aumento da precisão nas previsões de valores futuros.

A obtenção de valores da expressão gênica em tempos não estudados reduz custos relacionados com os procedimentos laboratoriais, os quais segundo Faceli, Carvalho e Souto (2005) são bastante onerosos e até limitantes para a implantação de projetos na área de *microarray*.

Este trabalho tem por objetivo avaliar qual a melhor alternativa, dentre os métodos de agrupamentos hierárquicos (Ward) e de otimização (Tocher), na formação de grupos homogêneos de séries de expressão gênica para posterior ajuste de modelos autorregressivos, AR(p), para dados em painel. Além disso, por meio deste ajuste, busca-se também realizar previsões da expressão gênica para observações futuras dentro de cada grupo formado e comparar com valores preditos obtidos por meio da análise individual de cada série de expressão.

2 MATERIAL E MÉTODOS

Os dados utilizados no presente estudo referem-se à expressão de genes que atuam sobre ciclo celular de *Saccharomyces cerevisiae* (ZHU et al., 2000). Os dados originais compreendem um experimento fatorial 2 x 2, sendo um fator a sincronização do ciclo celular por meio do componente alfa (sincronizado e não sincronizado) e o outro, diferentes cepas (selvagens e mutantes). Estes experimentos foram repetidos sequencialmente ao longo de 13 diferentes instantes de tempo equidistantes (0, 15, 30, ..., 180 minutos). Os experimentos em questão não possuem repetições, ou seja, os valores de *fold-change* ($\log_2$ da

razão de intensidade de luz emitida pelos genes do grupo tratado e do grupo controle) são provenientes de apenas um *slide* de *microarray*.

Todo o conjunto de dados utilizado está disponível no seguinte endereço eletrônico: <http://smd.stanford.edu/cgi-bin//publication/viewPublication.pl?pub_no=74>.

No presente trabalho foram usados apenas 10 tempos dos dados de células não sincronizadas, portanto, os valores de *fold-change* são provenientes da expressão de cepas mutantes (grupo tratado) em relação a cepas selvagens (grupo controle) dentro desta classe de células, em cada um dos tempos avaliados.

Inicialmente, foram considerados 3.607 genes, de forma que cada um deles apresentava 10 valores de *fold-change*. Com o intuito de realizar previsões para a última (10ª) observação, isto é, para o tempo 135 minutos, considerou-se nas análises apenas as observações referentes aos 9 primeiros tempos.

As estimativas dos parâmetros utilizadas como variáveis na análise de agrupamento foram obtidas por meio de ajustes individuais do modelo AR(p) para cada série de expressão, ou seja, neste ajuste prévio não foi considerada a teoria de dados em painel. Os modelos individuais foram ajustados tendo como variável resposta o valor de *fold-change*, considerando o seguinte modelo:

$$Y_j = \mu + \phi_1 Y_{j-1} + \phi_2 Y_{j-2} + ... + \phi_p Y_{j-p} + e_j,\ j=1,2,...,\ n_i\ . \qquad (1)$$

em que: Y_j é o valor do *fold-change* no tempo j, ϕ_k é o k-ésimo parâmetro de autorregressão; e_j é o termo de erro aleatório, $e_j \sim N(0,\sigma^2)$.

Devido ao pequeno número de observações de cada série (9 observações), ajustaram-se aos dados de *MTS* apenas os modelos autorregressivos de ordem $p \leq 5$. A escolha do melhor modelo foi realizada com base no critério de Schwartz, conhecido como critério de informação bayesiana

(BIC). De acordo com esse critério, menores valores de BIC refletem um melhor ajuste. Sua expressão é dada por:

$$BIC(M) = n\ln(L(y \mid \hat{\theta})) + M\ln(n),$$

em que: n é número de observações disponíveis para o ajuste, M é o número de parâmetros do modelo e $L(y \mid \hat{\theta})$ é o valor assumido pela função de verossimilhança quando utilizam-se as estimativas dos parâmetros do modelo.

Foram considerados para a análise de agrupamento somente os genes cujo BIC indicou os modelos autorregressivos de segunda ordem, AR(2), como sendo os mais plausíveis e que possuíam ambos os parâmetros autorregressivos significativos ($\alpha = 0{,}05$). Tal critério foi considerado com o objetivo de ilustrar a metodologia proposta tendo em vista a situação multiparamétrica mais simples, que se caracteriza pelo ajuste do modelo AR(2).

Posteriormente, realizou-se a análise de agrupamento considerando como variáveis no processo de agrupamento as estimativas dos parâmetros do modelo AR(2), ou seja, a média $\hat{\mu}_i$ (efeito do gene), $\hat{\phi}_{i1}$ e $\hat{\phi}_{i2}$ e a estimativa da variância do erro, $\hat{\sigma}_i^2$. Assim, o agrupamento foi realizado considerando quatro variáveis ($\hat{\mu}$, $\hat{\phi}_1, \hat{\phi}_2$ e $\hat{\sigma}^2$).

Para o agrupamento das séries de expressão gênica temporal utilizaram-se os métodos de Ward (1963) e Tocher (CRUZ; REGAZZI; CARNEIRO, 2004; FERREIRA, 2008; JOHNSON; WICHERN, 2007). No método de Ward os grupos são formados por meio da maximização da homogeneidade dentro dos mesmos, isto é, unem-se dois grupos A e B que minimizam o incremento na soma de quadrado do erro. Este incremento é definido como:

$$I_{AB} = \frac{n_A n_B}{n_A + n_B}(\bar{y}_A - \bar{y}_B)'(\bar{y}_A - \bar{y}_B),$$

em que: n_A: número de indivíduos que pertencem ao grupo A; n_B: número de indivíduos que pertencem ao grupo B; $\bar{y}_A$: vetor de média referente aos indivíduos que pertencem ao grupo A e; $\bar{y}_B$: média dos valores da variável Y dentro do grupo B.

Por outro lado, no método de Tocher, adota-se o critério de que a média das medidas de dissimilaridade dentro de cada grupo deve ser menor que as distâncias médias entre quaisquer outros grupos. A entrada de um indivíduo em um grupo sempre aumenta o valor médio da distância dentro do grupo. Assim, pode-se tomar a decisão de incluir o indivíduo em um grupo por meio da comparação entre o acréscimo no valor médio da distância dentro do grupo e um nível máximo permitido, que pode ser estabelecido arbitrariamente ou, adotado o valor máximo (θ) da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo para o seu grupo. Desta forma, a inclusão, ou não, do indivíduo k no grupo é, então, feita considerando:

- Se $\frac{d_{(grupo)k}}{n} \leq \theta$, inclui-se o indivíduo k no grupo;
- Se $\frac{d_{(grupo)k}}{n} > \theta$, o indivíduo k não é incluído no grupo.

em que: n representa o número de indivíduos que constitui o grupo original. Neste caso, a distância entre o indivíduo k e o grupo formado pelos indivíduos ij é dada por $d_{(ij)k} = d_{ik} + d_{jk}$ (CRUZ; REGAZZI; CARNEIRO, 2004). Uma característica interessante do método de Tocher é que ao final do processo de agrupamento o número de partições em que os indivíduos são alocados é conhecido automaticamente. Já nos métodos de agrupamento hierárquicos esse número não é fornecido, necessitando assim de um procedimento para obtenção do mesmo.

O número “ótimo” de grupos (partições), para o método de agrupamento hierárquico utilizado, foi obtido por meio do índice *RMSSTD (Root Mean Square Standard Deviation)* o qual é utilizado para calcular a homogeneidade dos agrupamentos (KHATTREE; NAIK, 2000). O cálculo do *RMSSTD*, para cada novo grupo formado é realizado por meio da seguinte expressão:

$$RMSSTD_{k} = \sqrt{\frac{SQ_1 + SQ_2 + ... + SQ_p}{gl_1 + gl_2 + ...gl_p}},$$

em que: $SQ_j = \sum_{i=1}^{n} \left(y_{ij} - \bar{y}_j\right)^2$ representa a soma de quadrado da j-ésima variável calculada considerando as n observações presentes em cada novo grupo k, ou seja, a cada novo grupo obtém-se um novo valor para o índice em questão; e gl_j representa o número de graus de liberdade referentes a j-ésima variável.

Para a obtenção do número “ótimo” de grupos o comportamento do *RMSSTD* em relação ao número de grupos foi modelado por um modelo exponencial, $RMSSTD = a(NG)^{-b}$, em que a e b são os parâmetros deste modelo, e NG corresponde ao número de grupos formados. Assim, como no trabalho de Cecon et al. (2008), o número “ótimo” de grupos foi determinado geometricamente por meio da interseção desta curva com uma reta, de forma que a maior distância entre elas corresponda ao ponto em questão (ponto de máxima curvatura).

Após a definição do melhor método de agrupamento para obtenção de grupos que contemplem a homogeneidade requerida para o estudo de dados em painel, realizou-se, para cada grupo formado, a análise considerando a estrutura de dados em painel de acordo com o seguinte modelo:

$$Y_{ij} = \mu + \phi_{i1} Y_{i(j-1)} + \phi_{i2} Y_{i(j-2)} + ... + \phi_{ip} Y_{i(j-p)} + e_{ij}, \quad i=1, ..., g \text{ e } j=1,2,..., n_i . \tag{2}$$

em que: Y_{ij} é o valor do *fold-change* do i-ésimo gene no tempo j, ϕ_{ik} é o k-ésimo parâmetro de autorregressão referente ao gene i; e_{ij} é o termo de erro aleatório, $e_{ij} \overset{i.i.d.}{\sim} N(0,\sigma^2)$.

O modelo (2) foi ajustado considerando a técnica de variáveis indicadoras por meio do PROC MODEL do *software* SAS® (SAS INSTITUTE INC., 2010), no qual é possível ajustar modelos lineares e não lineares com estrutura de erro autorregressivo.

Os intervalos de confiança foram obtidos de acordo com o exposto em Morettin e Toloi (2006):

$$\hat{Y}_{j+h} - \upsilon_\gamma \hat{\sigma}\left[1+\sum_{j=1}^{h-1}\psi_j^2\right]^{1/2} \le Y_{j+h} \le \hat{Y}_{j+h} + \upsilon_\gamma \hat{\sigma}\left[1+\sum_{j=1}^{h-1}\psi_j^2\right]^{1/2},$$

em que: $\hat{Y}_{j+h}$ é o valor predito da observação no tempo j+h, υ_γ é o γ-ésimo quantil da distribuição normal padrão e ψ_j^2 são os quadrados dos pesos de um filtro linear.

Os resultados obtidos pela análise de dados em painel foram comparados com as análises de séries individuais, ou seja, em que os parâmetros foram estimados de acordo com o modelo (1).

3 RESULTADOS E DISCUSSÃO

Dentre as 3.607 séries de expressão gênica, 222 apresentaram menores valores de BIC quando modeladas por processos autorregressivos de segunda ordem, AR(2). Destas 222 séries, 114 apresentaram ambos os coeficientes autorregressivos significativos ($\alpha = 0{,}05$) e, assim foram utilizadas na análise de agrupamento.

Após a obtenção das estimativas dos parâmetros individuais de cada série, foram realizadas as análises de agrupamento pelos métodos de Tocher e Ward.

Na Tabela 1 estão representados o número de genes, as médias e os desvios-padrão das estimativas dos parâmetros obtidos para cada grupo formado pelo método de Tocher tendo como medida de dissimilaridade o quadrado da distância euclidiana.

Por meio da Tabela 1 verifica-se a formação de 13 grupos nos quais a maioria de genes foram alocados no grupo 1.

Tabela 1 Número de genes, médias e desvios-padrão das estimativas dos parâmetros para cada grupo formado pelo método de Tocher, utilizando como medida de dissimilaridade o quadrado da distância euclidiana

Grupo	n° genes	$\overline{\hat{\mu}}$	$\overline{\hat{\phi}}_1$	$\overline{\hat{\phi}}_2$	$\overline{\hat{\sigma}}_e^2$
1	56	0,064	0,978	-0,717	0,041
		(0,112)	(0,145)	(0,097)	(0,038)
2	12	-0,106	1,266	-0,741	0,032
		(0,122)	(0,073)	(0,085)	(0,037)
3	15	-0,014	0,607	-0,699	0,026
		(0,099)	(0,052)	(0,114)	(0,017)
4	16	-0,072	-0,754	-0,655	0,014
		(0,097)	(0,101)	(0,097)	(0,012)
5	3	0,365	0,891	-0,803	0,076
		(0,031)	(0,082)	(0,047)	(0,052)
6	3	0,279	-0,805	-0,734	0,015
		(0,048)	(0,178)	(0,052)	(0,011)
7	2	0,106	1,434	-0,877	0,037
		(0,041)	(0,186)	(0,070)	(0,012)
8	2	-0,128	-0,917	-0,874	0,004
		(0,200)	(0,088)	(0,019)	(0,002)
9	1	0,115	-1,311	-0,866	0,002
		-	-	-	-
10	1	-0,021	-0,448	-0,929	0,011
		-	-	-	-
11	1	-0,101	0,839	-0,918	0,061
		-	-	-	-
12	1	0,573	1,003	-0,719	0,305
		-	-	-	-
13	1	0,480	1,291	-0,921	0,065
		-	-	-	-

Nota: Entre parênteses estão indicados os valores dos desvios-padrão.

Em relação ao método de WARD, a Figura 2 representa a determinação do número “ótimo” de grupos.

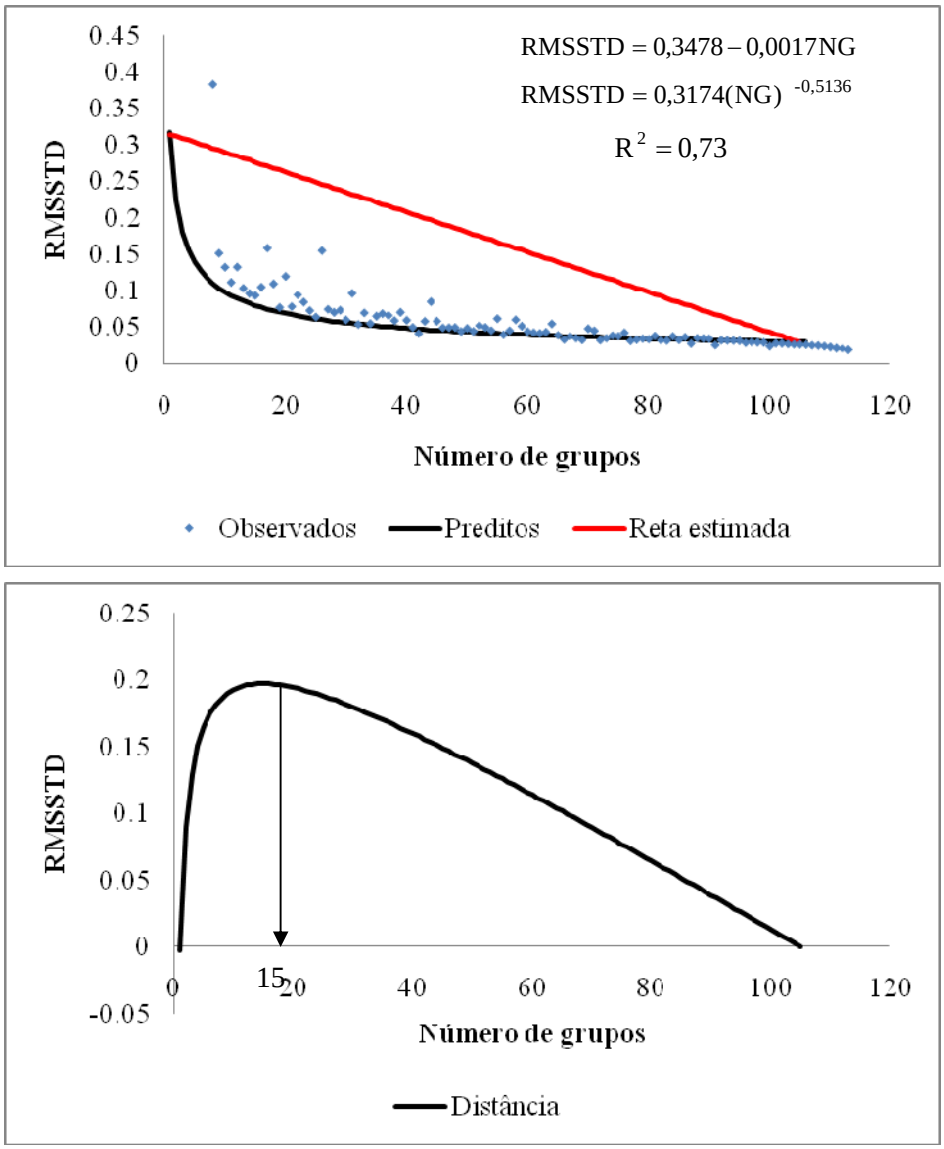

Figura 1 Determinação gráfica do número "ótimo" de grupos considerando as estimativas dos parâmetros do modelo AR(2), $\hat{\mu}_i$, $\hat{\phi}_{i1}$, $\hat{\phi}_{i2}$ e $\hat{\sigma}_i^2$

Verificou-se que o número "ótimo" de partições foi 15, quando o agrupamento foi realizado por meio do método de Ward utilizando como medida de dissimilaridade o quadrado da distância euclidiana (Figura 2).

Na Tabela 3 estão representados o número de genes, as médias e os desvios-padrão das estimativas dos parâmetros obtidos para cada grupo formado pelo método de Ward.

Tabela 2 Número de genes, médias e desvios-padrão das estimativas dos parâmetros para cada grupo formado pelo método de Ward utilizando como medida de dissimilaridade o quadrado da distância euclidiana

Grupo	n° de genes	$\overline{\hat{\mu}}$	$\overline{\hat{\phi}}_1$	$\overline{\hat{\phi}}_2$	$\overline{\hat{\sigma}}_e^2$
1	7	0,201	1,119	-0,843	0,038
		(0,062)	(0,062)	(0,041)	(0,014)
2	16	-0,043	0,937	-0,675	0,031
		(0,054)	(0,091)	(0,065)	(0,025)
3	15	0,103	0,809	-0,787	0,036
		(0,075)	(0,088)	(0,078)	(0,036)
4	9	-0,003	1,227	-0,678	0,032
		(0,077)	(0,049)	(0,049)	(0,032)
5	8	-0,187	1,204	-0,724	0,038
		(0,054)	(0,109)	(0,057)	(0,038)
6	8	0,035	0,712	-0,573	0,074
		(0,077)	(0,103)	(0,023)	(0,070)
7	5	0,316	0,883	-0,774	0,057
		(0,073)	(0,060)	(0,053)	(0,047)
8	6	0,046	1,331	-0,869	0,029
		(0,061)	(0,122)	(0,046)	(0,016)
9	12	-0,043	-0,801	-0,622	0,016
		(0,096)	(0,105)	(0,088)	(0,013)
10	7	0,132	1,018	-0,629	0,043
		(0,053)	(0,088)	(0,049)	(0,024)
11	11	-0,066	0,574	-0,768	0,023
		(0,082)	(0,039)	(0,090)	(0,018)
12	7	-0,129	-0,676	-0,814	0,008

Continua... Tabela 2

		(0,096)	(0,140)	(0,079)	(0,004)
13	3	0,279	-0,805	-0,734	0,015
		(0,048)	(0,178)	(0,052)	(0,011)
14	2	0,526	1,147	-0,820	0,185
		(0,066)	(0,204)	(0,143)	(0,170)
15	1	0,115	-1,311	-0,866	0,002
		-	-	-	-

Nota: Entre parênteses estão indicados os valores dos desvios-padrão.

Diante dos resultados apresentados, percebe-se que o método de agrupamento de Tocher tende a fornecer um menor número de grupos e, além disso, a maioria das séries de expressão gênica são alocadas no primeiro grupo (Tabelas 1 e 2). Essa característica observada no método de Tocher ocasiona, para o primeiro grupo formado, uma menor homogeneidade das estimativas dos parâmetros, μ, ϕ_1, ϕ_2 e σ_e^2, visto que os seus desvios-padrão são proporcionalmente maiores em relação aqueles apresentados para a maioria dos grupos formados pelo método de Ward.

Em face destes resultados, tendo em vista o objetivo deste estudo, que é a formação de grupos homogêneos a fim de se utilizar a metodologia de dados em painel para estudos de *MTS*, acredita-se que o método de Ward, seja a alternativa de agrupamento mais interessante.

Sob o ponto de vista biológico, um maior número de grupos e consequentemente grupos com menos genes, possibilita a realização de estudos de reação em cadeia da polimerase com transcrição reversa (RT-PCR), os quais têm por objetivo a análise da expressão de um ou poucos genes em um processo patológico ou em um processo fisiológico normal (LOPES, 2006). Além disso, esta técnica possibilita a validação de resultados obtidos por meio de estudos de *MTS*.

Uma vez escolhido o método de Ward ajustou-se, com base no modelo 2, o modelo AR(2) para dados em painel e obteve-se os valores preditos da expressão gênica em um tempo futuro (135 minutos). Assim, para fins de comparação entre os valores preditos e o verdadeiro valor da observação (excluído da análise) e buscando evitar a exaustividade de informações similares, são apresentados apenas resultados provenientes do primeiro grupo (ver Tabela 2) o qual consta de 7 genes. Nas Tabelas 3 e 4 estão representados, para cada gene do grupo 1, o verdadeiro valor da última observação (Y_{135}), sua estimativa ($\hat{Y}_{135}$), intervalos de confiança de 95%, erro quadrático médio de previsão (EQMP) e amplitude média dos intervalos de confiança (AM) com base nas estimativas dos parâmetros provenientes dos modelos 1 e 2, respectivamente.

Tabela 3 Verdadeiro valor da última observação (Y_{135}), seu valor predito ($\hat{Y}_{135}$), intervalos de confiança (95%), erro quadrático médio de previsão (EQMP) e amplitude média dos intervalos de confiança (AM) tendo em vista cada gene do grupo 1, formado pelo método de agrupamento de Ward (Modelo 1)

Gene	Y_{135}	Li	$\hat{Y}_{135}$	Ls
14	0,160	-0,242	0,188	0,617
20	-0,380	-1,498	-0,614	0,270
61	-0,385	-1,436	-0,614	0,207
65	-0,098	-0,745	-0,196	0,353
67	0,004	-0,494	0,148	0,789
80	0,129	-0,803	-0,185	0,433
105	0,397	-0,188	0,150	0,487
	EQMP		0,045	
	Amplitude média		1,232	

Nota: Li: limite inferior e Ls: limite superior.

Tabela 4 Verdadeiro valor da última observação (Y_{135}), seu valor predito ($\hat{Y}_{135}$), intervalos de confiança (95%), erro quadrático médio de previsão (EQMP) e amplitude média dos intervalos de confiança (AM) tendo em vista cada gene do grupo 1, formado pelo método de agrupamento de Ward (Modelo 2)

Gene	Y_{135}	Li	$\hat{Y}_{135}$	Ls
14	0,160	0,111	0,188	0,265
20	-0,380	-0,723	-0,614	-0,505
61	-0,385	-0,732	-0,614	-0,496
65	-0,098	-0,281	-0,196	-0,111
67	0,004	0,067	0,148	0,229
80	0,129	-0,275	-0,185	-0,095
105	0,397	0,064	0,150	0,436
	EQMP		0,045	
	AM		0,213	

Nota: Li: limite inferior e Ls: limite superior.

A porcentagem de intervalos de confiança que continham os verdadeiros valores de expressão gênica referentes ao tempo de 135 minutos foi de 100 e 86% para os modelos 1e 2, respectivamente.

Tanto o percentual de concordância entre os sinais dos verdadeiros valores da última observação (Y_{135}) e suas estimativas ($\hat{Y}_{135}$) (86%), quanto o valor do EQMP (0,045) foram igual para ambos os modelos ajustados (Tabelas 3 e 4). Devido ao uso da técnica de variáveis indicadoras na estimação do modelo 2 percebe-se que os valores preditos por ambos modelos ajustados são iguais. Por outro lado, visto que na estimação do erro são utilizadas todas as observações contidas no grupo, a amplitude média (AM=0,213) dos intervalos de confiança obtidos pela análise considerando a estrutura de dados em painel foi inferior aquele obtido pela análise individual (AM=1,232). Esse fato evidencia o aumento da precisão da estimativa do erro quando da utilização da estrutura de dados em painel.

Sob a perspectiva biológica, um alto percentual de concordância entre os sinais dos verdadeiros valores e suas estimativas indica que o modelo é capaz de classificar os genes como *up* ou *down regulated* de forma eficiente, ou seja, é possível prever em um tempo futuro se o gene se expressará mais no tratamento ou no controle. Além disso, a obtenção de valores preditos da expressão gênica reduz custos relacionados com os procedimentos laboratoriais, os quais muitas vezes são fatores limitantes para estudos de *MTS.*

4 CONCLUSÕES

A análise de agrupamento pelo método de Ward mostra-se mais apropriada na determinação de grupos que satisfaçam a pressuposição de homogeneidade requerida para a análise de dados em painel;

A análise de dados considerando uma estrutura de dados em painel apresenta-se mais eficiente em relação às análises individuais de cada série de expressão, uma vez que o valor de AM é inferior ao obtido pela análise individual de cada série.

REFERÊNCIAS

BAR-JOSEPH, Z. Analyzing time series gene expression data**. Bioinformatics**, Oxford, v. 20, p. 2493–2503, 2004.

BAR-JOSEPH, Z. et al. Continuous representations of time series gene expression data. **Journal of Computational Biology**, New York, v. 3, p. 341–356, 2003.

CECON, P. R. et al. Análise de medidas repetidas na avaliação de clones de café "Conilon". **Pesquisa agropecuária brasileira**, Brasília, v. 43, p. 1171-1176, 2008.

COSTA, I. G.; CARVALHO, F. A. T.; SOUTO, M. C. P. Comparative analysis of clustering methods for gene expression time course data. **Genetics and Molecular Biology**, Ribeirão Preto, v. 27, p. 623-631, 2004.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. 3. ed. Viçosa, MG: UFV, 2004. v. 1, 480 p.

EISEN, M. B. et al. Cluster analysis and display of genome-wide expression patterns. **Proceedings of the National Academy of Sciences of America,** Washington, v. 95, p. 14863-14868, 1998.

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

FERREIRA, D. F. **Estatística multivariada**. Lavras: UFLA, 2008. 662 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th ed. New Jersey: Prentice Hall, 2007. 800 p.

KHATTREE, R.; NAIK, D. **Multivariate data reduction and discrimination with SAS software**. Cary: SAS Institute, 2000. 574 p.

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

LOPES, A. C. **Diagnostico e tratamento**. São Paulo: Manole, 2006. 2112 p.

MORAIS, T. S. S. et al. Análise bayesiana de sensibilidade do modelo AR(1) para dados em painel: uma aplicação em dados temporais de microarrays. **Revista Brasileira de Biometria**, São Paulo, v. 28, p. 171-192, 2010.

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

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

RAMONI, M. F.; SEBASTIANI, P.; KOHANE, I. S. Cluster analysis of gene expression dynamics. **Proceedings of the National Academy of Sciences of America,** Washington, v. 99, p. 9121-9126, 2002.

SAS INSTITUTE INC. **SAS/STAT® 9.2**: user´s guide. 2nd ed. Cary, 2009.

SCHLIEP, A.; SCHONHUTH, A. STEINHOFF, C. Using hidden Markov models to analyze gene expression time course data. **Bioinformatics**, Oxford, v. 19, p.264-272, 2003.

SCHWARTZ, G. Estimating the dimension of a model. **Annals of Statistics**, Philadelphia, v. 6, p. 461-464, 1978.

SILVA, F. F. et al. Bayesian analysis of autoregressive panel data model: application in genetic evaluation of beef cattle. **Scientia Agrícola**, Piracicaba, v. 68, p. 237-245, 2011.

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v. 58, p. 236-244, 1963.

ZHU, G. et al. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. **Nature**, London, v. 406, p. 90-94, 2000.

APÊNDICE A - Procedimentos utilizados no *Software* SAS® (2010)

```
/*importação dos dados e combinação em um único arquivo*/
proc import datafile='C:\Moysés\Tese-MoysésUFLA\dadosSacchanomuces cerevisiae\000min.xls'
out=t000;
proc print;
run;
proc sort nodupkey NODUPRECS data=t000; by gene; run;
proc print data=t000;run;
.
.
.
proc import datafile='C:\Moysés\Tese-MoysésUFLA\dadosSacchanomuces cerevisiae\180min.xls'
out=t180;
proc print;
run;
proc sort nodupkey NODUPRECS data=t180; by gene; run;
proc print data=t180;run;
data fim; merge t000 (in=s1) t015 (in=s2) t030 (in=s3) t045 (in=s4) t060 (in=s5) t075 (in=s6) t090 (in=s7) t105 (in=s8) t120 (in=s9) t135 (in=s10) t150 (in=s11) t165 (in=s11) t180 (in=s12); if s1 & s2 & s3 & s4 & s5 & s6 & s7 & s8 & s9 & s10 & s11 & s12; by gene;run;
data fim1; set fim; run;
proc print data=fim1;run;

/*Preparação do banco de dados com 9 observações*/
```

```
proc transpose data=fim1 out=fim2; by gene;
 var  y000 y015 y030 y045 y060 y075 y090 y105 y120 ;
run;
proc print data=fim2; run;
data fim3; set fim2; rename COL1=y;
if _NAME_ = "y000" then do; t=0;output;end;
if _NAME_ = "y015" then do; t=15;output;end;
.
.
.
 if _NAME_ = "y120" then do; t=120;output;end;
run;
data fim4; set fim3; keep gene y t;
proc sort data=fim4;by gene;run;
```

```
/*Ajuste dos modelos autorregressivos e cálculo dos valores de BIC*/
*AR(1);
ods output EstSummaryStats=Log_ar1;
proc model data=fim4;by gene;
y=mu;
parms
mu=0;
 %ar(y,1);
fit y / fiml ;
run;
proc print data=Log_ar1;run;
data log_ar1_a; set Log_ar1; keep  bic_ar1 cValue2 gene ; if  cValue1=9;
bic_ar1=9*(cValue2)+ 1*log(9);
proc print data=log_ar1_a;run;
.
.
.
*AR(5);
ods output EstSummaryStats=Log_ar1;
proc model data=fim4;by gene;
y=mu;
parms
mu=0;
 %ar(y,5);
fit y / fiml ;
run;
 *proc print data=Log_ar3;run;
data log_ar5_a; set Log_ar5; keep  bic_ar5 cValue2 gene ; if  cValue1=9;
```

```
bic_ar5=9*(cValue2)+5*log(9);
proc print data=log_ar5_a;run;
data bic; merge  log_ar1_a log_ar2_a log_ar3_a log_ar4_a log_ar5_a; by gene;
run;
proc print data=bic;run;

/*Procedimento para obter as séries de expressão cujo BIC indicou o modelos
AR(2) como mais plausível*/
proc transpose data=bic out=teste1;
 var bic_ar1 bic_ar2 bic_ar3 bic_ar4 bic_ar5;by gene;
 run;
proc means min data=teste1;by gene;
var COL1;
output out=teste2 min=min;
run;
data teste3; merge teste1 teste2;by gene;run;
data teste4; set teste3; if COL1=min;run;
data ar2; set teste4;  if  _NAME_="bic_ar2"; run;

data fim41; merge fim4 (in=s1) ar2 (in=s2);if s1 & s2 ; keep  gene y; by
gene;run;
proc print data=fim41;run;;

/*Estimação dos parâmetros para entrada na análise de cluster*/
ods output  ResidSummary=MSE;
proc model data=fim41;by gene;
ods output MSE=fim16;
y=mu;
```

```
parms
mu=0;
 %ar(ynew,2);
fit ynew / outest=fim15;
run;
*proc print data=fim15;run;
data MSE1; set MSE; keep gene MS; run;
data fim16; merge fim15 MSE1; keep gene mu fi1 fi2 MS;run;

/*Análise de agrupamento*/
proc import datafile='C:\Moysés\Tese-MoysésUFLA\dadosSacchanomuces
cerevisiae\fim16.xls'
out=fim16;
run;
proc print data=fim16;run;
proc cluster  data=fim16 method=ward  rmsstd  out=tree_par;
id gene;
var fi1 fi2 ms mu;
run;
*proc print data=tree_par;run;

/*Cálculo do valor de RMSSTD e obtenção do número "ótimo" de grupos*/
/*Alguns procedimentos foram realizados com o auxílio do Microsoft Excel*/
data tree_par1; set tree_par; keep _NCL_  _RMSSTD_ ;
if  _RMSSTD_ ne 0;
run;
proc sort data=tree_par1;by _NCL_;run;
proc nlin data=tree_par1 method=dud;
```

```
parms a=0.3474  b=-0.00017;
model _rmsstd_=a*_ncl_**(-b);
run;
output out=tree_par2 predicted=yp;run;
proc print data=tree_par2;run;

/*Divisão dos grupos*/
proc tree data=tree_par nclusters=15 out=tree_par_3;
copy fi1 fi2 ms  mu; run;
data tree_par_31; set tree_par_3;rename _NAME_=gene;run;
proc sort data=tree_par_31; by  gene; run;
proc print data=tree_par_31;run;
proc sort data=tree_par_31; by cluster; run;

/*Obtenção das séries de expressão do primeiro grupo formado*/
data cluster1; set tree_par_31; keep gene;
if cluster=1;
run;
data cluster1fim ; merge fim1 (in=s1) cluster1(in=s2); if s1 & s2; by gene; run;

/*Análise de dados em painel*/
/*para o grupo composto por 7 genes*/

proc import datafile='C:\Moysés\Tese - Moysés UFLA\Pós-defesa\Análise
SAS Pós-defesa\grupo11.xls'
out=cluster1parcentrado;
proc print data=cluster1parcentrado;
run;
```

```
proc sort data=cluster1parcentrado;by gene;run;
proc print data=cluster1parcentrado;run;
proc transpose data=cluster1parcentrado out=teste1; by gene;
  var  y000       y015     y030     y045     y060     y075     y090     y105     y120;
run;
proc print data=teste1;run;

/*Obtenção das variável defasada*/
data ar1_0;set teste1; drop COL1;
 do _NAME_="y000";output;end; run;
data ar1_1; set ar1_0;
 do COL1=.;output;end; run;
proc sort nodupkey NODUPRECS data=ar1_1; by gene; run;
proc print data=ar1_1;run;
data ar1_2; set teste1; if _NAME_ ne "y120";
proc sort  data=ar1_2; by gene; run;
proc print data=ar1_2;run;
data teste2; set ar1_1 ar1_2; by gene;    rename COL1=y1;        drop _NAME_;
proc print data=teste2;run;
data teste3; set teste1; drop _NAME_; rename COL1=y;
proc print data=teste2;run;
data ar2_0;set teste1; drop COL1;
 do _NAME_="y01";output;end;run;
 proc sort nodupkey NODUPRECS data=ar2_0; by gene; run;
proc print data=ar2_0;run;
data ar2_1; set teste1; drop COL1;
do _NAME_="y02";output;end; run;
 proc sort nodupkey NODUPRECS data=ar2_1; by gene; run;
```

```
proc print data=ar2_1;run;
data ar2_2; set ar2_0 ar2_1;by gene;do COL1=.;output;end; run;
proc print data=ar2_2;run;
data ar2_3; set teste1; if _NAME_ = "y120" or _NAME_ = "y105" then delete;
proc sort  data=ar2_3; by gene; run;
proc print data=ar2_3;run;
data ar2_4; set ar2_2 ar2_3; by gene;    rename COL1=y2;        drop _NAME_;
proc print;run;
data final; merge teste2 teste3 ar2_4; drop _LABEL_;by gene; run;
proc print data=final;run;
data final_test;
 do y000=1 to 1;
.
.
.
 do y120=1 to 1;
output;
 end;
.
.
.

 end;
proc print data=final_test;run;

proc transpose data=final_test out=teste_fim;
 var  y000 y015 y030 y045   y060 y075 y090 y105 y120;
run;
```

```
proc print data=teste_fim;run;
data teste_fim1; set teste_fim;drop col1;
 do gene1= 1 to 7;output;end;run;
 proc print data=teste_fim1;run;
proc sort data=teste_fim1;by gene1;run;
proc print data=teste_fim1;run;
data final_ult; merge final teste_fim1;    drop gene1       ; run;
proc print data=final_ult;run;

/*Variáveis indicadoras*/
data final1; set final;
id1=(gene="YCR098C");
id2=(gene="YDL127W");
id3=(gene="YJL187C");
id4=(gene="YKL113C");
id5=(gene="YKR077W");
id6=(gene="YMR076C");
id7 =(gene="YOR255W");
run;
proc print data=final1;run;

/*Estimação do modelo*/
ods output ParameterEstimates=test_par_p;
proc model data=final1;
y =
id1*(u1 + a1*y1           +b1*y2)          +
.
.
```

```
.
id7*(u7 + a7*y1          +b7*y2);
parms
u1=0
.
.
.
u7=0
a1=0
.
.
.
a7=0
b1=0
.
.
.
b7=0;
fit y /;
run;
proc print data= test_par_p;run;
```

ARTIGO 2 Agrupamento de séries de expressão gênica por meio de estimativas provenientes de análise bayesiana do modelo autorregressivo para dados em painel

RESUMO

Foi proposta uma nova metodologia para o agrupamento de genes com padrões de expressões gênicas similares, com base nas estimativas dos parâmetros provenientes da análise bayesiana do modelo autorregressivo de ordem p, AR(p), para dados em painel. Além disso, objetivou-se realizar previsões baseadas em distribuições preditivas para valores de expressões gênicas em tempos futuros. Os dados utilizados referem-se à expressão de genes que atuam sobre ciclo celular de *Saccharomyces cerevisiae*. Tais dados correspondem a 114 genes, sendo que, cada um deles apresentava 10 valores de *fold-change* (medida da expressão) ao longo do tempo (0, 15, 30, ..., 135 min). A metodologia proposta foi capaz de agrupar genes que compartilham de padrões de expressão similares. Além disso, foi possível a obtenção de boas estimativas para os valores preditos.

Palavras-chave: *Microarray time series*. Previsão. *Saccharomyces cerevisiae*.

Grouping temporal gene expression by using estimates from bayesian analysis of autoregressive model for panel data

ABSTRACT

A new methodology was developed for the clustering of genes with similar expression patterns, based on parameter estimates from the Bayesian analysis of the autoregressive model of order p, AR (p) for panel data. Additionally, the purpose was to make predictions based on predictive distributions of gene expression values in the future. Data of the expression of genes that control the cell cycle of *Saccharomyces cerevisiae* were used, corresponding to 114 genes, of which each had 10 fold-change values (measure of expression) over time (0, 15, 30, ..., 135 min.) The proposed methodology was able to group genes with similar expression patterns and secondly obtained good estimates for the predicted values.

Index terms: *Microarray time series*. Forecasting. *Saccharomyces cerevisiae*.

1 INTRODUÇÃO

Uma das abordagens mais importantes na ciência genômica é a análise de experimentos de expressão gênica avaliada ao longo do tempo, também conhecida como *Microarray Time Series* (*MTS*). De acordo com Kim e Kim (2008) a observação temporal da expressão gênica possibilita ao pesquisador caracterizar o gene por meio de seu padrão longitudinal de expressão. Devido ao grande número de genes avaliados numa análise de *MTS*, agrupar os genes que compartilham padrões similares é o primeiro passo para o entendimento de redes biológicas complexas, as quais possibilitam atribuir funções para os genes analisados (SCHLIEP; SCHONHUTH; STEINHOFF, 2003).

Segundo Schliep, Schonhuth e Steinhoff (2003), as metodologias para a análise de dados *MTS* podem ser divididas em duas classes. Na primeira, as observações da expressão em cada tempo são consideradas variáveis independentes, e assim métodos usuais como os de agrupamento hierárquico (EISEN et al., 1998) e o de otimização conhecido como k-médias (TAVAZOIE et al., 1999) podem ser diretamente utilizados para agrupar genes com padrão de expressão semelhantes.

Na segunda classe, o agrupamento baseia-se no ajuste de modelos específicos, portanto é considerada mais interessante sob o ponto de vista estatístico e biológico, uma vez que devido a utilização de tais modelos, a dependência temporal entre as observações pode ser levada em consideração no processo de agrupamento de genes. Dentre os métodos pertencentes a esta segunda classe, destacam-se os de agrupamentos baseados na dinâmica do padrão de expressão gênica (RAMONI; SEBASTIANI; KOHANE (2002); nos modelos de Markov oculto (SCHILIEP et al., 2003) e no agrupamento por meio de representações contínuas do tipo B-splines (BAR-JOSEPH et al., 2003). Porém, apesar destes métodos serem úteis, segundo Bar-Joseph (2004) os

mesmos não são adequados para experimentos relativamente pequenos, com menos de 10 observações temporais por gene.

De acordo com Ernst, Nau e Bar-Joseph (2005), os quais examinaram o banco de dados de Stanford (*Stanford Microarray Database - SMD*), mais de 80% de todas as séries continham menos de 8 observações temporais. Este fato destaca outra característica dos estudos de *MTS* além da grande quantidade de genes avaliados, que é o pequeno número de medidas de expressão temporal por gene.

De acordo com o exposto, uma metodologia bastante usada na área de séries temporais quando se tem um grande número de séries com poucas observações é a análise de dados em painel. Esta análise proporciona um aumento na precisão das estimativas dos parâmetros de interesse em relação a análises individuais de cada série, uma vez que há um aumento no número de observações devido à combinação das várias observações temporais de todos os indivíduos.

Tendo em vista as vantagens relatadas dos métodos de agrupamento baseados em modelos e a dificuldade de aplicá-los a séries com poucas observações, acredita-se que uma nova metodologia que considere modelos de séries temporais sob o enfoque de dados em painel seja eficiente para a análise de dados *MTS*. Tal metodologia possibilitaria descrever o padrão da expressão gênica por meio de dependências temporais, por exemplo, mediante ajuste de modelos autorregressivos, e simultaneamente agrupar genes cujas estimativas dos parâmetros dos modelos sejam similares. Além disso, ao se usar os referidos modelos também seria possível realizar previsões da expressão gênica para tempos não observados, fato este que implicaria em uma redução de tempo e custo para pesquisas envolvendo análise de dados *MTS*.

A análise de dados em painel sob o enfoque de modelos de séries temporais assume que os coeficientes de cada série sejam considerados amostras

de uma mesma distribuição, portanto tem-se naturalmente a pressuposição de uma distribuição para os parâmetros do modelo, a qual sob o ponto de vista bayesiano caracteriza-se como a distribuição a *priori*. Dessa forma, segundo Liu e Tiao (1980), Morais et al. (2010) e Silva et al. (2011) a inferência bayesiana mostra-se mais simples e adequada ao se trabalhar com modelos autorregressivos para dados em painel.

Diante do exposto, o presente trabalho tem como principal objetivo propor uma metodologia para o agrupamento de genes com padrões de expressões gênicas similares baseadas nas estimativas dos parâmetros provenientes da análise bayesiana do modelo autorregressivo de ordem p, AR(p), para dados em painel. Além disso, propõe-se também realizar previsões baseadas em distribuições preditivas para valores de expressões gênicas em tempos futuros.

2 MATERIAL E MÉTODOS

2.1 Descrição dos dados *MTS*

Os dados utilizados para a aplicação do método proposto referem-se à expressão de genes que atuam sobre ciclo celular de *Saccharomyces cerevisiae* (ZHU et al., 2000).

Os dados originais compreendem um experimento fatorial 2 x 2, sendo um fator a sincronização do ciclo celular por meio do componente alfa e o outro, diferentes cepas (selvagens e mutantes). Estes experimentos foram repetidos sequencialmente ao longo de 13 diferentes instantes de tempo (0, 15, 30, ..., 180 minutos). Os experimentos em questão não possuem repetições, ou seja, os valores de *fold-change* ($\log_2$ da intensidade de luz emitida pelos genes do grupo

tratado e do grupo controle), variável resposta (Y_{ij}), são provenientes de apenas um *slide* de *microarray*.

No presente trabalho foram usados apenas 10 tempos dos dados de células não sincronizadas, portanto, os valores de *fold-change* são provenientes da expressão de cepas mutantes (grupo tratado) em relação a cepas selvagens (grupo controle) dentro desta classe de células, em cada um dos tempos avaliados.

Inicialmente, foram considerados 3.607 genes, de forma que cada um deles apresentava 10 valores de *fold-change*. Com o intuito de realizar previsões para a última (10ª) observação, isto é, tempo 135 minutos, considerou-se nas análises apenas as observações referentes aos 9 primeiros tempos.

Com o objetivo de ilustrar a metodologia proposta tendo em vista a situação multiparamétrica mais simples, que se caracteriza pelo ajuste do modelo AR(2), foram considerados para a análise de agrupamento 114 genes cujo BIC indicou, dentre os modelos ajustados ($p \leq 5$), os modelos autorregressivos de segunda ordem, AR(2), como sendo os mais plausíveis e que possuíam ambos os parâmetros significativamente diferente de zero ($\alpha = 0{,}05$).

Todo o conjunto de dados completo está disponível no seguinte endereço eletrônico:<http://smd.stanford.edu/cgi-bin//publication/viewPublication.pl?pub_no=74>.

2.2 Análise bayesiana do modelo AR(p) para dados em painel

O modelo autorregressivo de ordem p, AR(p), para dados em painel, em que p é o número de parâmetros do modelo, está representado na equação 1:

$$Y_{it} = \mu_i + \phi_{i1} Y_{i(t-1)} + \phi_{i2} Y_{i(t-2)} + \ldots + \phi_{ip} Y_{i(t-p)} + e_{it} \qquad (1)$$

$$Y_{it} = \mu_i + \sum_{j=1}^{p} \phi_{i(t-j)} Y_{i(t-j)} + e_{it}$$

em que: i=1,2, ...,m; j=1,2,...,p; t=1,2,..., n_i; μ_i é a média do processo referente ao individuo i; Y_{it} é o valor atual de um processo estocástico, referente ao individuo i, $\phi_{i1}, \phi_{i2}, \ldots, \phi_{ip}$ são os parâmetros referente ao modelo para o i-ésimo indivíduo, denominados parâmetros de autorregressão; e_{it} é o resíduo associado ao modelo, também denominado de ruído branco, $e_{it} \overset{iid}{\sim} N(0, \sigma_e^2)$.

A função de verossimilhança, considerando $n = n_1 = n_2 = \ldots = n_m$, de acordo como o modelo apresentado (1), condicionadas as mp primeiras observações, é dada por:

$$L\left(\mathbf{Y} \mid \mathbf{\Phi}, \sigma_e^2\right) \propto \sigma_e^{2\,-\left(\frac{m(n-p)}{2}\right)} \exp\left\{ -\frac{1}{2\sigma_e^2} \sum_{i=1}^{m} \sum_{t=p+1}^{n} \left(Y_{it} - \mu_i - \sum_{j=1}^{p} \phi_{ij} Y_{i(t-j)} \right)^2 \right\}$$

Reescrevendo a função de verossimilhança em forma matricial para todos os indivíduos, tem-se:

$$L\left(\mathbf{Y} \mid \mathbf{\Phi}, \sigma_e^2\right) \propto \sigma_e^{2\,\frac{-m(n-p)}{2}} \exp\left\{ -\frac{1}{2\sigma_e^2} \left(\mathbf{Y} - \mathbf{X\Phi}\right)' \left(\mathbf{Y} - \mathbf{X\Phi}\right) \right\},$$

em que: $\mathbf{Y} = \left[y_{1(p+1)}, y_{1(p+2)}, \ldots, y_{1(n)}, y_{2(p+1)}, \ldots, y_{2(n)}, y_{m(p+1)}, \ldots, y_{m(n)} \right]'$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 & 0 & 0 \\ 0 & \mathbf{X}_2 & 0 & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{X}_m \end{bmatrix}_{m(n-p)\times m(p+1)},$$

$$\mathbf{X}_i = \begin{bmatrix} 1 & y_{i(p)} & \cdots & y_{i(1)} \\ 1 & y_{i(p+1)} & \cdots & y_{i(2)} \\ \vdots & & \ddots & \vdots \\ 1 & y_{i(n-1)} & \cdots & y_{i(n-p)} \end{bmatrix}_{(n-p)\times(p+1)} \quad \text{e}$$

$$\mathbf{\Phi} = \left[\mu_1, \phi_{11}, \phi_{12}, \ldots, \phi_{1p}, \mu_1, \phi_{21}, \ldots, \phi_{2p}, \ldots, \mu_m, \phi_{m1}, \ldots, \phi_{mp}\right]' \in R^{m(p+1)}.$$

Os vetores $\mathbf{\Phi}$ e $\mathbf{Y}$, considerados nas expressões apresentam, respectivamente, as dimensões m(p+1) x 1 e m(n-p) x 1(SILVA et al., 2011).

De acordo com a metodologia bayesiana, para a estimação dos parâmetros do modelo AR(p), faz-se necessário atribuir distribuições a *priori* para os parâmetros de interesse $\mathbf{\Phi}$ e σ_e^2. Neste estudo, da mesma forma que no trabalho de Silva et al. (2008b) considerou-se a *priori* hierárquica Normal multivariada - Gama Inversa, representada como segue:

$$P\left(\mathbf{\Phi}, \sigma_e^2\right) = P\left(\mathbf{\Phi} \mid \sigma_e^2\right) P\left(\sigma_e^2\right)$$

em que: $\left(\mathbf{\Phi} \mid \sigma_e^2\right) \sim N_{m(p+1)}\left(\boldsymbol{\mu}, \sigma_e^2 \mathbf{I}\right)$ e $\sigma_e^2 \sim GI(\alpha, \beta)$ (Gama Inversa), em que $\mathbf{I}$ é uma matriz identidade de ordem m(p+1) x m(p+1). Assim, tem-se:

$$P\left(\mathbf{\Phi} \mid \sigma_e^2\right) \propto \sigma_e^{2^{-\left(\frac{m(p+1)}{2}\right)}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[\left(\mathbf{\Phi}-\boldsymbol{\mu}\right)' \mathbf{I} \left(\mathbf{\Phi}-\boldsymbol{\mu}\right)\right]\right\} \text{ e}$$

$$P\left(\sigma_e^2\right) \propto \sigma_e^{2^{-(\alpha+1)}} \exp\left\{-\frac{\beta}{\sigma_e^2}\right\}.$$

Portanto, a distribuição conjunta a *priori*, $P\left(\mathbf{\Phi}, \sigma_e^2\right)$, é dada por:

$$P\left(\boldsymbol{\Phi},\sigma_e^2\right)\propto \sigma_e^{2\,-\left(\frac{m(p+1)+2\alpha}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma_e^2}\left[2\beta+\left(\boldsymbol{\Phi}-\boldsymbol{\mu}\right)'\mathbf{I}\left(\boldsymbol{\Phi}-\boldsymbol{\mu}\right)\right]\right\}.$$

Os componentes $\boldsymbol{\mu}$, $\mathbf{I}$ (matriz identidade), α e β são denominados hiperparâmetros, e representam os parâmetros das distribuições a *priori* dos parâmetros do modelo considerado.

Combinando a função de verossimilhança, $L\left(\mathbf{Y}\mid\boldsymbol{\Phi},\sigma_e^2\right)$, com a distribuição a *priori*, $P\left(\boldsymbol{\Phi},\sigma_e^2\right)$, obtém-se, via Teorema de Bayes, a distribuição conjunta a *posteriori*:

$$P\left(\boldsymbol{\Phi},\sigma_e^2\mid\mathbf{Y}\right)\propto L\left(\mathbf{Y}\mid\boldsymbol{\Phi},\sigma_e^2\right)P\left(\boldsymbol{\Phi},\sigma_e^2\right),$$

$$P\left(\boldsymbol{\Phi},\sigma_e^2\mid\mathbf{Y}\right)\propto \sigma_e^{2\,\frac{-m(n-p)}{2}}\exp\left\{-\frac{1}{2\sigma_e^2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\Phi}\right)'\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\Phi}\right)\right\}\times$$

$$\sigma_e^{2\,-\left(\frac{m(p+1)+2\alpha}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma_e^2}\left[2\beta+\left(\boldsymbol{\Phi}-\boldsymbol{\mu}\right)'\mathbf{I}^{-1}\left(\boldsymbol{\Phi}-\boldsymbol{\mu}\right)\right]\right\},$$

$$P\left(\boldsymbol{\Phi},\sigma_e^2\mid\mathbf{Y}\right)\propto \sigma_e^{2\,-\left(\frac{m(n+1)+2\alpha}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma_e^2}\left[2D+\left(\boldsymbol{\Phi}-\hat{\boldsymbol{\Phi}}_B\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Phi}-\hat{\boldsymbol{\Phi}}_B\right)\right]\right\},$$

em que:

$$D=\beta+\frac{\left(\mathbf{Y}'\mathbf{Y}+\boldsymbol{\mu}'\mathbf{I}^{-1}\boldsymbol{\mu}\right)-\left(\mathbf{X}'\mathbf{Y}+\mathbf{I}^{-1}\boldsymbol{\mu}\right)'\left(\mathbf{X}'\mathbf{X}+\mathbf{I}^{-1}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}+\mathbf{I}^{-1}\boldsymbol{\mu}\right)}{2},$$

$$\hat{\boldsymbol{\Phi}}_B=\left(\mathbf{X}'\mathbf{X}+\mathbf{I}^{-1}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}+\mathbf{I}^{-1}\boldsymbol{\mu}\right)\text{ e }\ \boldsymbol{\Sigma}=\mathbf{X}'\mathbf{X}+\mathbf{I}^{-1}.$$

Para fazer inferências sobre os parâmetros de interesse é necessário obter suas distribuições marginais a *posteriori*. Estas distribuições são obtidas por meio da integração da distribuição a *posteriori* em relação a todos os parâmetros, exceto o de interesse.

Na maioria dos casos, essas integrais são complexas e não apresentam soluções exatas. Para contornar este problema utilizou-se algoritmos *MCMC* (*Markov Chain Monte Carlo*), como o Gibbs *sampl*er (GELFAND; SMITH, 1990) e/ou Metropolis-Hastings (HASTINGS, 1973). Foram obtidas as distribuições condicionais completas a *posteriori*, com o intuito de gerar valores indiretamente das distribuições marginais a *posteriori* por meio da teoria das cadeias de Markov.

Assim, após a obtenção da distribuição conjunta a *posteriori* faz-se necessário apresentar as distribuições condicionais completas a *posteriori* para $\boldsymbol{\Phi}$ e σ_e^2, sendo estas respectivamente:

$$\boldsymbol{\Phi} \mid \sigma_e^2, \mathbf{Y} \sim \mathrm{N}_{m(p+1)}\left(\hat{\boldsymbol{\Phi}}_B, \sigma_e^2 \boldsymbol{\Sigma}\right),$$

$$\sigma_e^2 \mid \boldsymbol{\Phi}, \mathbf{Y} \sim \mathrm{GI}\left(\frac{m(n+1)+2\alpha}{2}, D + \frac{1}{2}\left(\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}_B\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}_B\right)\right).$$

Pode-se perceber que as distribuições condicionais completas para os parâmetros $\boldsymbol{\Phi}$ e σ_e^2, são dadas respectivamente por uma distribuição normal multivariada e por uma distribuição gama-inversa e, portanto, passível ao uso do Gibbs Sampler.

O algoritmo[2] Gibbs Sampler foi implementado matricialmente no *software* estatístico R (R DEVELOPMENT CORE TEAM, 2011), sendo as funções *mnormt* (*multivariate Normal and t-Student distributions*) e *rinvgamma* (*inverse Gamma distribution)* utilizadas, respectivamente, para geração de números aleatórios das distribuições normal multivariada e gama inversa.

O algoritmo foi executado considerando uma cadeia de 10.000 iterações, de forma que as 2.000 primeiras foram eliminadas para o aquecimento da cadeia

[2] O algoritmo está apresentado no Apêndice A.

(*burn-in*). Para avaliar a convergência foi utilizado o critério de Raftery e Lewis (1992) mediante o pacote *Bayesian Output Analysis* (*BOA*) do R.

2.3 Formação de grupos homogêneos para análise de dados em painel

A análise de dados em painel fundamentada no modelo 1 pressupõe que as séries temporais de todos os indivíduos sejam homogêneos, pois ao contrário não seria possível assumir que os coeficientes sejam amostras de uma mesma distribuição. Assim, ao se ajustar modelos autorregressivos para dados em painel é necessário assumir certa homogeneidade entre as séries dos diferentes indivíduos (LIU; TIAO, 1980; MORAIS et al., 2010; SILVA et al., 2011).

Na análise de dados *MTS*, devido à grande quantidade de genes envolvidos, esta pressuposição não pode ser explicitamente assumida, fazendo-se necessário a determinação de grupos de genes homogêneos para a aplicação do método baseado na análise de dados em painel. Tais grupos foram obtidos por meio de um processo iterativo no qual inicialmente consideraram-se todos os genes como sendo provenientes de uma mesma população, o que permitiu a aplicação da análise bayesiana do modelo AR(p) para dados em painel e, consequentemente, a obtenção das estimativas dos parâmetros de cada série.

Em seguida, estas estimativas foram utilizadas como variáveis de entrada na análise de agrupamento, e para cada grupo formado ajustou-se novamente o modelo AR(p) para dados em painel. Este procedimento resultou em novas estimativas para os parâmetros, as quais foram mais uma vez submetidas à análise de agrupamento. Dessa forma, deu-se início a um processo iterativo, repetido até que o número de grupos (k) e os indivíduos pertencentes a eles não apresentassem mais alterações.

Na Figura 1 está representado o esquema do procedimento descrito para genes cujas séries de expressão foram modelados por um AR(2) para dados em painel.

Ao final do referido processo, obteve-se como resultado os grupos de séries similares de expressão gênica, de acordo com as estimativas dos parâmetros do modelo AR(p) para dados em painel, possibilitando a realização de previsões para valores futuros com base em distribuições preditivas específicas para cada grupo.

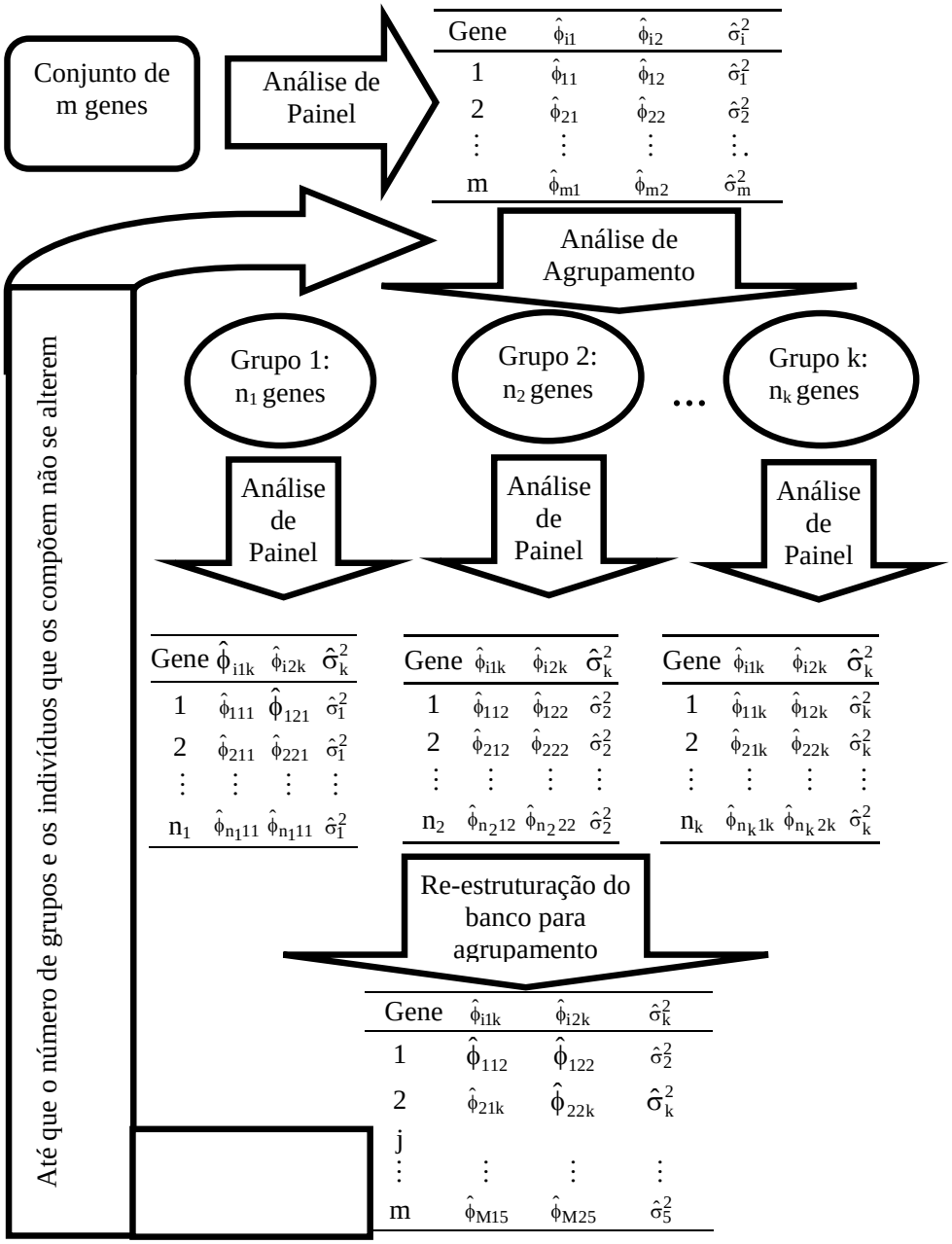

Figura 1 Esquema de análise para genes que são descritos por um modelo AR(2)

2.3.1 Método de agrupamento de Ward

Para a implementação do processo iterativo apresentado em 2.2 utilizou-se o método hierárquico de Ward (1963). Neste método os grupos são formados por meio da maximização da homogeneidade dentro dos grupos, isto é, unem-se dois grupos A e B que minimizam o incremento na soma do quadrado do erro. Este incremento é definido como:

$$I_{AB} = \frac{n_A n_B}{n_A + n_B}\left(\bar{y}_A - \bar{y}_B\right)'\left(\bar{y}_A - \bar{y}_B\right).$$

Como os métodos de agrupamento hierárquico não fornecem o número final de grupos, utilizou-se o método proposto por Mojema (1977). Esse método determina o número de grupos, k, que otimiza a qualidade do ajuste do agrupamento aos dados (FERREIRA, 2009). Deve-se escolher o número de grupos dado pelo primeiro estágio do dendrograma no qual:

$$\alpha_j > \bar{\alpha} + \psi S_\alpha ,$$

em que: j=1,2,...,n; α_j é o valor da distância para o estágio de junção correspondente a n-j+1 grupos; $\bar{\alpha}$ e S_α são: a média e o desvio padrão dos $\alpha's$, respectivamente e; ψ é uma constante que de acordo com Milligan e Cooper (1995) deve assumir valor 1,25.

2.3.2 Distribuição preditiva sob o enfoque de painel

Após a formação dos grupos utilizou-se a teoria de distribuição preditiva descrita por Heckman e Leamer (2001), que é dada por:

$$\mathbf{Y}_{(n+1)}^{(q)} \mid \mathbf{Y} \sim N_m\left(\mathbf{X}\hat{\mathbf{\Phi}}^{(q)}, \hat{\sigma}_e^{2(q)}\mathbf{I}\right),$$

em que: **I** é uma matriz identidade de ordem m(p+1) x m(p+1).

Desta forma, o conjunto de valores gerados para esta distribuição normal multivariada, proveniente da q-ésima iteração do algoritmo Gibbs Sampler, constituem a distribuição preditiva para um dado futuro, cuja estimativa do valor predito, $\hat{\mathbf{Y}}_{(n+1)}$, é representada pela média da distribuição $P(\mathbf{Y}_{(\mathbf{n+1})} \mid \mathbf{Y})$.

Para avaliar a capacidade preditiva do modelo AR(p) para dados em painel ajustado para cada grupo obtido de acordo com o item 2.1.1, utilizou-se o recurso apresentado por Liu e Tiao (1980), que consiste na remoção da última observação de cada série. Assim, para cada grupo, os parâmetros dos modelos foram estimados sem a presença destas observações, as quais foram preditas pela metodologia apresentada no presente item.

Uma forma prática de se avaliar a eficiência destas predições é verificar se os intervalos de credibilidade obtidos continham os verdadeiros valores excluídos da análise. Assim, foram calculadas as percentagens de intervalos de credibilidade que continham estes valores de expressão gênica referentes ao último tempo, de forma que quanto maior tal porcentagem maior a eficiência preditiva.

3 RESULTADOS E DISCUSSÃO

Dentre as 3.607 séries de expressão gênica, 222 apresentaram menores valores de BIC quando modeladas por processos autorregressivos de segunda ordem, AR(2). Destas 222 séries, 114 apresentaram ambos os coeficientes autorregressivos significativamente diferentes de zero ($\alpha = 0{,}05$) e, assim foram utilizadas na análise de agrupamento.

Para o agrupamento das 114 séries (padrões de expressão gênica) utilizou-se um processador Intel Core 2 Duo E7500 2.0 GHz com 4 GB de memória RAM. O algoritmo, implementado no *software* R, teve um tempo de execução de 24’46’’. Após a execução do algoritmo as séries de expressão

foram particionadas em 5 grupos distintos. Os números de séries de expressão gênica em cada grupo foram respectivamente, 23, 32, 15, 24 e 23 para os grupos 1, 2, 3, 4 e 5.

Em relação à convergência, para todas as cadeias simuladas, o fator de dependência de Raftery e Lewis (1992) forneceu valores menores que 5. Além disso, de acordo com esse mesmo critério os valores máximos de *burn-in*, *thinning* e do total de iterações para a obtenção de convergência foram inferiores a 500, 2 e 4.000 respectivamente.

As Figuras 2 a 7 representam os *Box-plot* para os valores de expressão gênica compreendidos entre os tempos de 0 a 135 minutos, para cada grupo de genes pela metodologia proposta. Na Figura 2 está representada a situação em que todos os genes são considerados pertencentes a um único grupo (primeira fase do processo iterativo), enquanto que as demais figuras (Figuras 3 a 7), apresentam os *Box-plot* dos 5 grupos formados ao final da execução do método proposto.

Percebe-se que a ausência do agrupamento de genes (Figura 2) proporciona uma maior variabilidade das observações em cada tempo avaliado, uma vez que o número de observações, respectivamente abaixo e acima dos quantis 2,5% e 97,5%, é consideravelmente maior em relação aos grupos formados (Figuras 3 a 7). Este resultado evidencia a necessidade de se utilizar a análise de agrupamento para a obtenção de grupos de genes homogêneos, uma vez que esta é uma pressuposição do modelo AR(p) para dados em painel.

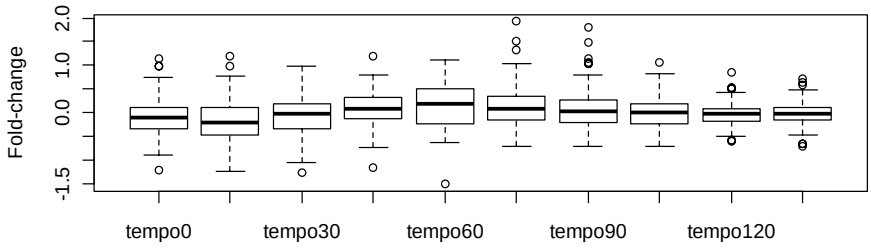

Figura 2 *Box-plot* referente aos valores de expressão ao longo do tempo considerando todos os genes como pertencentes a um único grupo

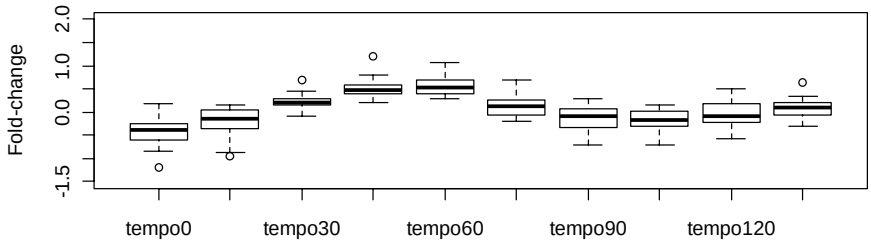

Figura 3 *Box-plot* referente aos valores de expressão ao longo do tempo do grupo 1

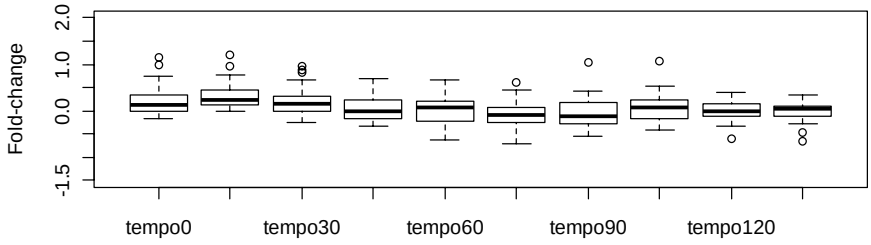

Figura 4 *Box-plot* referente aos valores de expressão ao longo do tempo do grupo 2

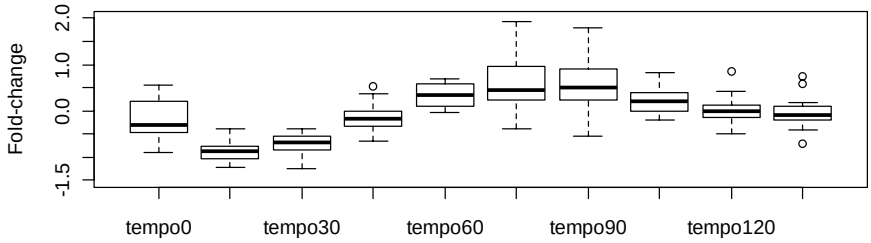

Figura 5 *Box-plot* referente aos valores de expressão ao longo do tempo do grupo 3

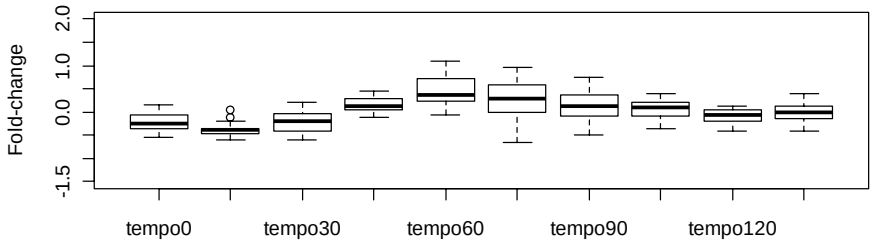

Figura 6 *Box-plot* referente aos valores de expressão ao longo do tempo do grupo 4

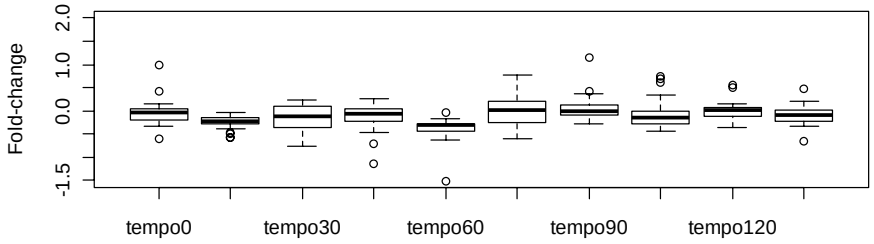

Figura 7 *Box-plot* referente aos valores de expressão ao longo do tempo do grupo 5

Nas Figuras 7 e 8 estão representadas as séries temporais de expressão gênica pertencentes a cada grupo formado e, as séries médias dos valores observados em cada tempo dentro de cada grupo. De maneira geral, pode-se observar que os 5 grupos formados possuem padrões de expressão bem distintos. Dentre as diversas diferenças, pode-se notar que os genes que compõem os grupos 1 e 2 possuem comportamento médio inverso durante o ciclo celular, ou seja, os valores observados de *fold-change* são respectivamente, para os grupos 1 e 2, negativos e positivos até um dado tempo e, após esses tempos os mesmos invertem de sinais (Figuras 8B e 8D). Os genes que pertencem ao grupo 5 em geral se expressam mais no controle (cepas selvagens) durante o ciclo celular de *saccharomyces.*

Nas Figuras 8A, 8C, 8E, 9A e 9C estão representadas as séries de expressão temporais de todos os genes pertencentes a cada grupo formado. Percebe-se que existem diferenças quanto ao número de genes que compõem cada grupo. Esta diferença pode estar associada ao número de funções às quais os genes estão associados, isto é, grupos com um maior número de genes podem estar associados a um maior número de funções durante o ciclo celular. Já os grupos que contenham um menor número de genes podem estar associados a funções mais específicas e, em estudos mais específicos podem ser de maior interesse para estudos de RT-PCR, os quais são utilizados para a validação de resultados obtidos em um estudos de *MTS*.

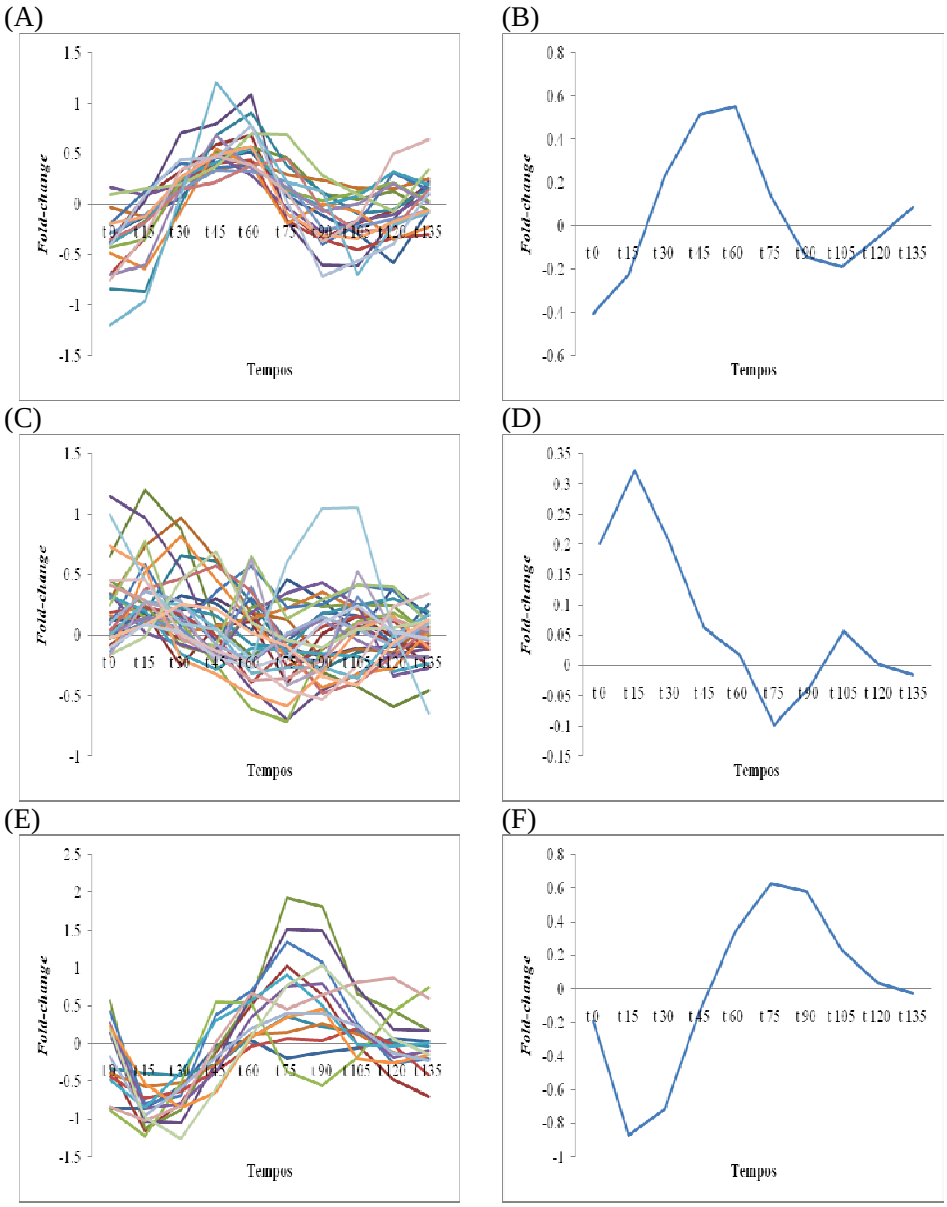

Figura 8 Séries de expressão de três grupos encontrados pela metodologia proposta. (A) Padrões de expressão do grupo 1; (B) Padrão médio de expressão do grupo 1; (C) Padrões de expressão do grupo 2; (D) Padrão médio de expressão do grupo 2; (E) Padrões de expressão do grupo 3; (F) Padrão médio de expressão do grupo 3

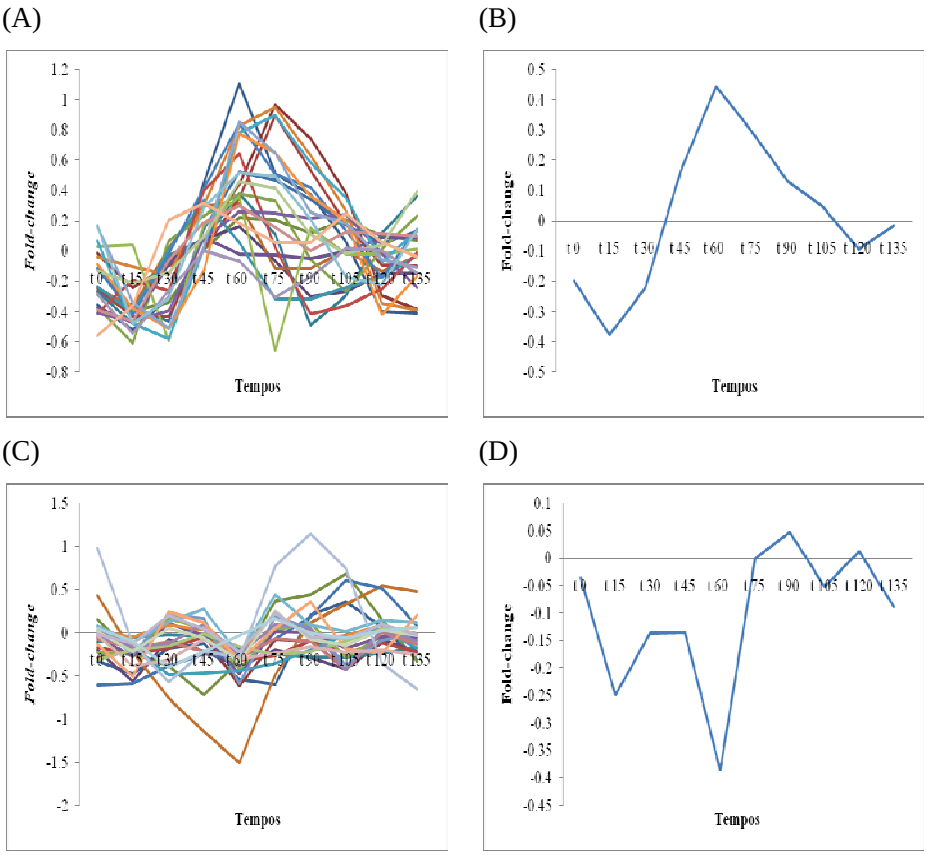

Figura 9 Séries de expressão de dois grupos encontrados pela metodologia proposta. (A) Padrões de expressão do grupo 4; (B) Padrão médio de expressão do grupo 4; (C) Padrões de expressão do grupo 5; (D) Padrão médio de expressão do grupo 5

Uma vez determinado o número de partições e as séries que as compõem, obtiveram-se os valores preditos da expressão gênica em um tempo futuro. Assim, para fins de comparação entre os valores preditos e o verdadeiro valor da observação (excluído da análise) e buscando evitar a exaustividade de informações similares, são apresentados apenas resultados provenientes do primeiro grupo o qual consta de 20 genes.

A Tabela 1 representa os verdadeiros valores da última observação, isto é, o valor de *fold-change* observado a 135 minutos (Y_{135}) de cada gene, seu valor predito ($\hat{Y}_{135}$), os respectivos intervalos de credibilidade de 90% para cada grupo formado e o erro quadrático médio de previsão.

Tabela 1 Verdadeiros valores da última observação (Y_{135}) de cada gene, suas estimativas ($\hat{Y}_{135}$) e os limites inferior (LI) e superior (LS) dos intervalos de credibilidade de 90%

Gene	Y_{135}	LI	$\hat{Y}_{135}$	LS
1	-0,051	-2,707	-0,009	2,687
3	0,002	-2,655	0,044	2,726
12	0,220	-2,672	0,226	3,131
13	-0,192	-2,897	-0,163	2,617
14	0,182	-2,533	0,170	2,912
30	0,201	-2,777	0,197	3,091
35	0,028	-2,763	0,070	2,898
40	-0,069	-3,116	-0,111	2,959
46	0,205	-2,653	0,202	3,027
51	0,022	-2,724	0,057	2,805
55	0,162	-2,667	0,209	3,080
56	0,127	-2,636	0,140	2,932
62	0,029	-2,715	0,016	2,764
71	0,072	-2,664	0,067	2,774
79	0,271	-2,596	0,251	2,944
83	0,116	-2,703	0,098	2,856
88	-0,079	-3,136	-0,052	2,953
94	-0,012	-2,814	-0,029	2,786
99	-0,191	-3,042	-0,182	2,666
108	-0,050	-2,874	-0,042	2,830
	EQMP		0,001	

De acordo com a Tabela 6 pode-se afirmar que a metodologia utilizada para realizar previsões de dados futuros individuais com base na obtenção das distribuições preditivas foi eficiente, uma vez que o valor do Erro Quadrático Médio de Previsão (0,001) foi muito pequeno. Além disso, verifica-se que ambas as porcentagens de concordância entre os sinais dos verdadeiros valores da última observação (Y_{135}) e suas estimativas ($\hat{Y}_{135}$) e de intervalos de credibilidade que continham os verdadeiros valores de expressão gênica referentes ao tempo 135 foram de 100%.

Sob o aspecto biológico, 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, Carvalho e Souto (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 (MORAIS et al., 2010). Em relação à concordância entre os sinais dos verdadeiros valores e suas estimativas tem-se como principal incremento a possibilidade de ser classificar os genes, em um tempo futuro, como *up* ou *down regulated*, ou seja, é possível prever se o gene se expressará mais no tratamento ou no controle.

No trabalho de Morais et al. (2010) é apresentada uma relação de diversos autores que também avaliaram a capacidade preditiva de modelos autorregressivos sob enfoque bayesiano mediante intervalos de credibilidade a *posteriori*. Como exemplos pode-se citar o trabalho de Alba (1993), que ao simular quatro séries independentes sob um modelo autorregressivo de ordem quatro, AR (4), obteve 75% de eficiência na predição de um dado futuro, e Silva

et al. (2008a), 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. Esses resultados corroboram com a afirmação de que a metodologia utilizada para realizar previsões foi eficiente, uma vez que o resultado obtido neste estudo (71%) não se distancia dos demais.

Na Tabela 2 estão representadas as estimativas obtidas para a variância do erro ($\hat{\sigma}_e^2$) dentro de cada grupo formado.

Tabela 2 Estimativas da variância do erro ($\hat{\sigma}_e^2$), intervalos de credibilidade de 90% (Li: limite inferior e Ls: limite superior) para cada grupo formado

Grupo	$\hat{\sigma}_e^2$	Li	LS
1	0,517	0,456	0,583
2	0,373	0,335	0,412
3	0,621	0,541	0,706
4	0,458	0,407	0,514
5	0,459	0,406	0,513

De acordo com a Tabela 2, percebe-se que a variância do erro foi maior no grupo 3 seguida do grupo 1. Esse resultado indica que as boas estimativas para os valores preditos obtidos para o primeiro grupo sejam também verificadas para os demais.

Em estudos futuros pretende-se avaliar a eficiência da técnica proposta em relação às outras metodologias encontradas na literatura e, verificar até que ponto a utilização de valores preditos não altera o resultado obtido por um processo de agrupamento. Além disso, objetiva-se também avaliar se diferentes formas de obtenção de grupos similares, seja por métodos baseados em identidade de modelos, em otimização, ou até mesmo a utilização de diferentes

critérios para o obtenção do número “ótimo” de grupos para os métodos de agrupamento hierárquicos, incorporarão alguma melhoria ao método proposto.

4 CONCLUSÕES

1. A metodologia proposta foi capaz de agrupar genes que compartilham de padrões de expressão similar.

2. A metodologia utilizada para realizar previsões de dados futuros individuais com base na obtenção das distribuições preditivas foi eficiente, uma vez que o valor do EQMP (0,001) foi muito pequeno.

REFERÊNCIAS

ALBA, E. Constrained forecasting in autoregressive time series models: a bayesian analysis. **International Journal of Forecasting**, New York, v. 9, p. 95-108, 1993.

BAR-JOSEPH, Z. Analyzing time series gene expression data. **Bioinformatics**, Oxford, v. 20, p. 2493-2503, 2004.

BAR-JOSEPH, Z. et al. Continuous representations of time series gene expression data. **Journal of Computational Biology**, New York, v. 3, p. 341–356, 2003.

EISEN, M. B. et al. Cluster analysis and display of genome-wide expression patterns. **Proceedings of the National Academy of Sciences of America,** Washington, v. 95, p. 14863-14868, 1998.

ERNST, J.; NAU, G. J.; BAR-JOSEPH, Z. Clustering short time series gene expression data. **Bioinformatics**, Oxford, v. 21, p. 159-168, 2005.

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

FERREIRA, D. F. **Estatística multivariada**. Lavras: UFLA, 2008. 662 p.

GELFAND, A. E.; SMITH, A. F. M. sampling based approaches to calculating marginal densities**, Journal of the American Statistical Association**, Alexandria**,**
v. 85, p. 398-409, 1990.

HASTINGS, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. **Biometrika**, London, v. 57, p. 97–109, 1970.

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

KIM, J.; KIM, H. Clustering of change patterns using Fourier coefficients. **Bioinformatics**, Oxford, v. 24, p. 184-191, 2008.

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

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, Colorado Springs, v. 50, p. 159-179, 1985.

MOJEMA, R. Hierarquical grouping methods and stopping rules: na evaluation. **Computer Journal**, Trier, v. 20, p. 359-363, 1977.

MORAIS, T. S. S. et al. Análise bayesiana de sensibilidade do modelo AR(1) para dados em painel: uma aplicação em dados temporais de microarrays. **Revista Brasileira de Biometria**, São Paulo, v. 28, p. 10, 2010.

RAFTERY, A. E.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J. M. et al. **Bayesian statistics 4.** Oxford: Oxford University, 1992. p. 763-773.

RAMONI, M. F.; SEBASTIANI, P.; KOHANE, I.S. Cluster analysis of gene expression dynamics. **Proceedings of the National Academy of Sciences of America,** Washington, v. 99, p. 9121-9126, 2002.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2008. Disponível em: <http://www.R- project.org>. Acesso em: 21 jan. 2011.

SCHLIEP, A.; SCHONHUTH, A. STEINHOFF, C. Using hidden Markov models to analyze gene expression time course data. **Bioinformatics**, Oxford, v. 19, p. 264-272, 2003.

SCHWARTZ, G. Estimating the dimension of a model. **Annals of Statistics,** Philadelphia, v. 6, p. 461-464, 1978.

SILVA, F. F. et al. Bayesian analysis of autoregressive panel data model: application in genetic evaluation of beef cattle. **Scientia Agrícola**, Piracicaba, v. 68, p. 237-245, 2011.

SILVA, F. F. et al. Comparação bayesiana de modelos de previsão de diferenças esperadas nas progênies no melhoramento genético de gado Nelore. **Pesquisa Agropecuária Brasileira**, Brasília, p. 37-45, 2008a.

SILVA, F. F. et al. Previsão bayesiana de valores genéticos de touros por meio do modelo autorregressivo para dados em painel. **Arquivo Brasileiro de Medicina Veterinária *e* Zootecnia**, Belo Horizonte, p. 1166-1173, 2008b.

TAVAZOIE, S. et al. Systematic determination of genetic network architecture. **Nature Genetics,** New York, v. 22, p. 281-285, 1999.

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v. 58, p. 236-244, 1963.

ZHU, G. et al. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth**. Nature**, London, v. 406, p. 90-94, 2000.

APÊNDICE A - Códigos de programação no *software* R.

```
###Leitura dos dados e instalação dos pacotes necessários###############
library(MCMCpack)
library(mnormt)
dados<-read.table("dados_painel_completo_AR(2).txt",header=T)
grupos<-rep(1,nrow(dados))
dados2<-cbind(grupos,dados)
#####################################################################

##########Função Amostrador de Gibbs para painel##################

# A função deve receber:
# dados1: matriz de dados;
#iter: número de iterações
#nlag: ordem do modelo autorregressivo
#burnin: número de iterações excluídas para o aquecimento da cadeia.

gibbspanel<-function(dados1,iter,nlag,burnin)
{
 m<-nrow(dados1)
 p<-nlag
############Chutes iniciais e vetores estimados#######################
 fi_est<-matrix(0,nrow=iter,(nrow(dados1)*p)+nrow(dados1))
 sigma_e<-matrix(0, nrow=iter,1)
 fi_est[1,]<- t(t(fi_est[1,]))
 sigma_e[1]<-1
 dados_agrup<-matrix(0,nrow=nrow(dados1),p+2)
```

```
############valores que variam a cada iteração#########################
 m=nrow(dados1)
 n=ncol(dados1)-2
#####Definindo os valores para os hiper-parâmetros (mu, P, alpha, beta)########
 P=diag(1,(m*p)+m,(m*p)+m)
 mu=matrix(0,(m*p)+m,1)
 alpha1=50
 beta1=49
#########################GIBBS###################################
 y<-dados1[,3:ncol(dados1)] #vetor de observações
 b<-matrix(t(y[,p:(n-1)]),m*(n-p),1)
 c<-matrix(t(y[,1:(n-p)]),m*(n-p),1)
 X_aux<-matrix(0,m*(n-p),m*(p+1))
 M<-matrix(1,n-p,3) #3= número de parâmetros
 X_auxi<-kronecker(diag(1, m), M)
 Xi<-matrix(cbind(1,b,c),m*(n-p),m*(p+1))
 X<-Xi*X_auxi
 Y1<-matrix(t(y[,(p+1):n]),m*(n-p),1)
 sigma_M<-t(X)%*%X+solve(P) #matriz covariância da condicional de fi
 fiB<-solve(sigma_M)%*%(t(X)%*%Y1+P%*%mu) #média da condicional de fi
 D<-beta1 +0.5*((t(Y1)%*%Y1+t(mu)%*%solve(P)%*%mu)
           -t((t(X)%*%Y1+solve(P)%*%mu))%*%
            solve(sigma_M)%*%(t(X)%*%Y1+solve(P)%*%mu))
 shape1<-0.5*((m*(n+1))+(2*alpha1))
 for(iter in 2:iter)
  {
```

```
  fi_est[iter,]<-mvrnorm(n=1, fiB, sigma_e[iter-1]*sigma_M)
  sigma_e[iter,]<-rinvgamma(1,shape1,D+0.5*(t(t(t(fi_est[iter,]))-
fiB)%*%solve(sigma_M)%*%(t(t(fi_est[iter,]))-fiB)))
  }
 a<<-fi_est
 b<<-sigma_e
 h<-as.data.frame(fi_est)
 media1<-mean(h[burnin:iter,])
 media<-matrix(media1,1,(m*p)+m)
 precisao<-mean(sigma_e[burnin:iter,])
 sigma_e1_grupo<-rep(precisao,nrow(dados1))
 dados_agrup[,4]<-t(t(sigma_e1_grupo))
 dados_agrup[,1]<-t(media[1,c(seq(1,3*m,3))])
 dados_agrup[,2]<-t(media[1,c(seq(2,(3*m)-1,3))])
 dados_agrup[,3]<-t(media[1,c(seq(3,(3*m),3))])
 dados_agrup1<-cbind(dados1[,2],dados1[,1],dados_agrup)
 return(list(dados_agrup1=dados_agrup1))
}
```

####################Fim da Função################################

###Função agrupamento e determinação do número "ótimo" de grupos###

```
agrupamento<-function(dados_agr,dados1)
{
 mdist<-dist(dados[,3:6], method="euclidean")
 agrup<-hclust(mdist^2,method ="ward")
 mojema=mean(agrup$height)+1.25*sd(agrup$height)
 k=length(agrup$height[agrup$height>mojema]) + 1
```

```
 grupos<-cutree(agrup, k=k)
 grupos1<<-as.data.frame(grupos)
 dados1<-cbind(grupos,dados1[,-1] )
 return(list(dados1=dados1,grupos=grupos))
}
```

####################Fim da Função##################################

########Processo iterativo para o agrupamento de genes################

```
date()
grupo_teste<-matrix(0,100,ncol=nrow(dados1))
i<-2
repeat
{
  x<-gibbspanel(dados1,10000,2,2000)
  agr<-agrupamento(x$dados_agrup1,dados1)
  grupo_teste[i,]<-t(as.matrix(agr$grupos))
  x1<-by(agr$dados1,agr$dados1[,1], function(x) gibbspanel(x,10000,2,2000))
  x2<-do.call("rbind",x1)
  x2<-x2[order(x2[,1]),]
  grupos1<-t(t(agr$grupo))
  dados1<-cbind(grupos1,dados1[,-1])
  if(sum(grupo_teste[i,]-grupo_teste[i-1,])==0)
  break
  i<-i+1
}
date()
```

####################Fim do processo##################################

######################Previsão###################################

```
iter=10000
fi_prev<-matrix(0,nrow=iter,nrow(dados1))
y<-dados1[,3:ncol(dados1)]
m=nrow(dados1)
p=2
q<-cbind(rep(1,nrow(dados1)),y[,12],y[,11])
aa<-matrix(c(1,0,0,0,1,0,0,0,1),3,m*(p+1))
qq<-q%*%aa
M1<-matrix(c(1,1,1),1,3)
A<-kronecker(diag(1, m),M1)
X1<-A*qq
I<-kronecker(diag(1,m),1)
for(w in 1:iter)
{
fi_prev[w,]<-mvrnorm(n=1, X1%*% t(t(a[w,])), b[w]*I)
}
previsao<-matrix(0,nrow=nrow(dados1),1)
for( d in 1:nrow(dados1))
{
previsao[d,]<-mean(fi_prev[2000:10000,d])
}
```

####################Fim do cálculo###################################

###########Cálculo dos Intervalos de Credibilidade ##################

```
ICP<-matrix(0,nrow=ncol(fi_prev),ncol=2)
for(l in 1:ncol(fi_prev))
{
ICP[l,]<(cbind(quantile(fi_prev[2000:10000,l],0.05),quantile(fi_prev[2000:1000
0,l],0.95)))
}
```

####################Fim do cálculo################################

####################Salvar saídas################################

```
write.table(conv, file = "./nome_do_arquivo_de_interesse.txt")
```

########################Fim ################################

CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo principal propor uma metodologia para o agrupamento de genes com padrões de expressões gênicas similares baseadas nas estimativas dos parâmetros provenientes de uma análise bayesiana do modelo autorregressivo de ordem p, AR(p), para dados em painel. Além disso, buscou-se apresentar os fundamentos teóricos a serem utilizados como ferramenta para a implementação desta proposta.

Na primeira parte foi apresentada a teoria de modelos autorregressivos para dados em painel, uma breve introdução sobre inferência bayesiana e os principais métodos de simulação *MCMC*. Além disso, foi apresentada a abordagem bayesiana do modelo AR(p), para dados em painel. Posteriormente, foram apresentados alguns critérios para a seleção de modelos, métodos de agrupamento hierárquicos e de otimização e a determinação do número "ótimo" de grupos para os métodos hierárquicos. Finalmente, realizou-se uma breve introdução sobre dados de *MTS* (*Microarray Time Series*).

A segunda parte desta tese é composta por dois artigos. A análise realizada, nos dois artigos, considera dados referentes à expressão de genes que atuam sobre ciclo celular de *Saccharomyces cerevisiae*. Tais dados correspondem a 106 genes, sendo que, cada um deles apresentava 13 valores de *fold-change* (medida da expressão) ao longo do tempo (0, 15, 30, ..., 165, 180 minutos).

No primeiro artigo objetivou-se verificar a viabilidade da utilização de métodos de agrupamentos, hierárquico (WARD) e de otimização (Tocher), na formação de grupos homogêneos de séries de expressão gênica para posterior ajuste de modelos autorregressivos, AR(p), para dados em painel. Além disso, buscou-se também ajustar o modelo AR(p) para dados em painel, possibilitando assim a realização de previsões da expressão gênica dentro de cada grupo

formado. De acordo com os resultados, o método de WARD mostrou-se mais apropriado na obtenção de grupos de genes cujas séries foram consideradas homogêneas. Posteriormente, ajustou-se o modelo AR(2) para dados em painel e obteve-se com sucesso a predição da expressão gênica em tempos futuros.

No segundo artigo, apresentou-se uma nova metodologia para o agrupamento de genes baseada nas estimativas dos parâmetros provenientes da análise bayesiana do modelo autorregressivo de ordem p, AR(p), para dados em painel. Além disso, mostrou-se que é possível obter boas estimativas para os valores futuros por meio de previsões baseadas em distribuições preditivas. A metodologia proposta mostrou-se capaz e eficiente para agrupar genes que compartilham de padrões de expressão similares.

Acredita-se que a abordagem do segundo artigo, quando comparada àquela utilizada no primeiro artigo, seja mais eficiente na obtenção de grupos homogêneos de séries de expressão gênica temporal, uma vez que, a análise bayesiana proporciona uma maior precisão das estimativas dos parâmetros de interesse em relação a análises individuais de cada série e, consequentemente proporciona uma maior acurácia ao processo de agrupamento.

Como perspectivas de estudos futuros, pretendem-se avaliar a eficiência da técnica proposta em relação às outras metodologias encontradas na literatura e, verificar até que ponto a utilização de valores preditos não altera o resultado obtido por um processo de agrupamento. Além disso, objetiva-se também avaliar se diferentes formas de obtenção de grupos similares, seja por métodos baseados em identidade de modelos, em otimização, ou até mesmo a utilização de diferentes critérios para o obtenção do número “ótimo” de grupos para os métodos de agrupamento hierárquicos, incorporarão alguma melhoria ao método proposto.