Métodos Estatísticos Aplicados à Análise da Expressão Gênica

Erlandson Ferreira Saraiva

Orientador: Prof. Dr. Luís Aparecido Milan

Co-orientadora: Prof[a]. Dra. Teresa Cristina M. Dias

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos

Fevereiro de 2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S243me

Saraiva, Erlandson Ferreira.
Métodos estatísticos aplicados à análise da expressão gênica / Erlandson Ferreira Saraiva. -- São Carlos : UFSCar, 2006.
136 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Estatística matemática. 2. Expressão gênica. 3. Inferência bayesiana. 4. Seleção de modelos. 5. Processos de Dirichlet. I. Título.

CDD: 519.5 (20ª)

Agradeço,

à minha família pelo apoio e incentivo em todos os momentos,

à minha noiva Sandra pelo apoio e compreensão,

ao professor Dr. Luís Aparecido Milan pela orientação, pelas idéias e principalmente pelo exemplo de dedicação e disciplina de trabalho,

à professora Dra. Teresa Cristina M. Dias pela co-orientação e sugestões dadas no exame de qualificação,

ao professor Dr. José Galvão Leite pelas valiosas sugestões e esclarecimentos sobre os aspectos teóricos relacionados aos métodos estatísticos apresentados nesta dissertação,

à CAPES (Coordenação de Aperfeiçoamente de Pessoal de Nível Superior) pelo apoio financeiro.

Resumo

A tecnologia dos arranjos de DNA (DNA-array) é uma ferramenta utilizada para identificar e comparar níveis de expressão de um grande número de genes ou fragmentos de genes simultaneamente, em condições diferentes. Com esta comparação, é possível determinar possíveis genes causadores de doenças de origem genética (por exemplo, o câncer). Nestes experimentos, grandes quantidades de dados numéricos (relacionados às medidas de níveis de expressão dos genes) são gerados e métodos estatísticos são importantes para análise dos dados, com objetivo de identificar os genes que apresentam evidências para níveis de expressão diferentes. O objetivo de nossa pesquisa é comparar o desempenho e desenvolver métodos estatísticos, capazes de identificar genes que apresentam evidências para níveis de expressão diferentes, quando comparamos situações de interesse (tratamentos) com uma situação de controle. Para isto, descrevemos o teste t, proposto por Baldi e Long (2001) e propomos três métodos para identificar genes com evidências para níveis de expressão diferentes. O primeiro método proposto é baseado na utilização da inferência bayesiana paramétrica e dos métodos de seleção de modelos, fator de Bayes e DIC; o segundo método é baseado na inferência bayesiana semi-paramétrica conhecida como modelo de misturas de processos Dirichlet; e o terceiro método é baseado na utilização de um modelo com mistura infinita de distribuições, que aplicado à análise da expressão gênica determina grupos de níveis de expressão gênica similares, baseados nos efeitos de tratamento.

Palavras-chave: Arranjos de DNA, teste t, Inferência bayesiana, *Monte Carlo*, *Monte Carlo Markov Chain* - MCMC, *Gibbs Sampling*, Seleção de modelos, Fator de Bayes, DIC, Inferência bayesiana não paramétrica, Processo Dirichlet, Modelo de misturas de processos Dirichlet, Modelo com mistura de distribuições.

Abstract

The technology of the DNA-Arrays is a tool used to identify and to compare levels of expression of a great number of genes or fragments of genes, in different conditions. With this comparison, it is possible to identify genes possibly causing illnesses of genetic origin (cancer for example). Great amounts of numerical data (related the measures of levels of expression of the genes) are generated and statistical methods are important for analysis of this data with objective to identify the genes that present evidences for different levels of expression. The objective of our research is to develop and to describe methods statistical, capable of identifing genes that present evidences for different levels of expression. We describe the test t, considered for Baldi and Long (2001) and consider three others methods. The first method considered is based on the use of parametric Bayes inference and the methods for selection of models, Bayes factor and DIC; the second method is based an semi-parametric bayesian inference, model of mixtures of Dirichlet processes. The third method is based on the use of a model with infinite mixtures of distributions that applied the analysis of the genica expression determines groups of similar levels of expression.

Sumário

Capítulo 1

Introdução

Com o desenvolvimento da genética foram criados os termos transcriptoma, que representa o conjunto completo dos transcritos (RNA's) e proteoma, que representa o conjunto completo das proteínas.

À medida que mais genes vão sendo conhecidos, através dos projetos genomas, tem-se a possibilidade de passar para fases de análises seguintes: saber quando e onde estes genes são expressos, ou seja, o funcionamento do genoma, genoma funcional.

Segundo Felix *et al.*, (2002), "O fluxo de informação gênica do DNA nos cromossomos (genoma) até o proteoma, é intermediado pelo conjunto das moléculas de RNA (transcriptoma). Assim, a concentração relativa de transcritos de um determinado gene em uma célula é um indicativo do quanto esse gene está sendo expresso, isto é, do quanto a célula está investindo do seu maquinário bioquímico para produzir a proteína codificada pelo gene".

Com isso, pesquisadores de diversas áreas (biólogos, estatísticos, engenheiros, químicos, físicos, matemáticos, etc...) voltaram suas atenções ao desenvolvimento de tecnologias, visando medir a concentração relativa dos transcritos (RNA's) dos genes em células e tecidos. Uma das principais ferramentas para este tipo de estudo são os arranjos de DNA[1].

Os arranjos de DNA são lâminas (ou suportes), comumente de vidro ou náilon, que

[1]Existem diversos termos empregados para descrever os arranjos de DNA: náilon *array*, *filter array*, *high density membranes* e *macroarray* para os arranjos em suporte de náilon; e *glass array*, DNA *chips*, *biochips* e *microarray* para os arranjos em suporte de vidro (ver Felix *et al.*, 2002).

permitem medir os níveis expressão de uma grande quantidade de genes ou fragmentos de genes, ao mesmo tempo. Para isto, são fixados na(s) lâmina(s), de forma ordenada, seqüências completas ou parciais de genes conhecidos. Fixados os genes na(s) lâmina(s), produzem-se sondas de cDNA[2] com alguma marcação (radioativa ou fluorescente). Por exemplo, marca-se o cDNA de uma situação (por exemplo, células normais) com um marcador que reflete a luz vermelha, e o cDNA da outra situação (células doentes) com um marcador que reflete a luz verde (ou vice-versa). Misturam-se as sondas na(s) lâmina(s), se as sondas utilizadas possuírem seqüências complementares aos fixados na(s) lâmina(s), então vão se hibridar[3], permitindo sua identificação (ver Figura 1).

O resultado do experimento são duas imagens com pontos iluminados, uma de verde e a outra de vermelho (ver Figura 1). A imagem dos pontos é salva e processada por computadores e *softwares* específicos, onde cada *spot* ou "ponto" do arranjo recebe valores numéricos (ver Wu, 2001).

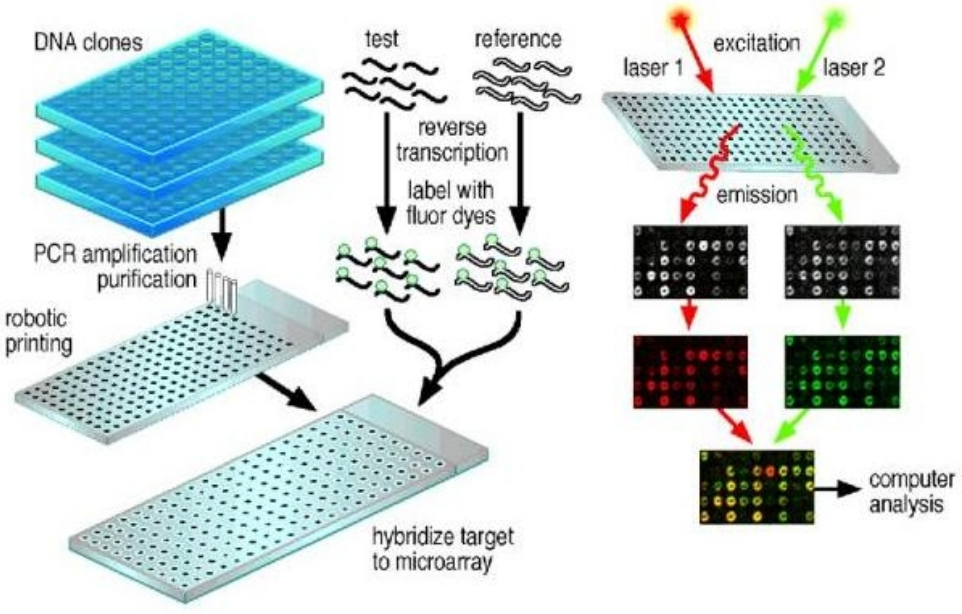

Figura 1: Experimento com arranjos de DNA.

Fonte: www.orfe.princeton.edu/research/microarray.jpg

[2]DNAs complementares: cópias, se possível, de todos os RNAs mensageiros de uma linhagem celular ou órgão na forma de DNA.

[3]A hibridação entre as moléculas de DNA, que estão na lâmina, com as da população de cDNA (derivado de mRNA extraído das células em estudo) é feita por complementaridade que as fitas de DNA apresentam se forem homólogas (de mesma origem).

Se as duas lâminas forem sobrepostas, obtém-se a última lâmina da Figura 1. Os pontos onde se observa uma coloração esverdeada, indicam que existe uma quantidade maior de mRNA[4] correspondentes, por exemplo, às células doentes em relação a células normais. Em outras palavras, o gene em questão pode estar mais expresso na célula doente, o que pode ter alguma implicação na origem e/ou evolução de uma doença. O raciocínio inverso é aplicado aos pontos do arranjo, em que se observa uma coloração avermelhada (ver Carneiro *et al.*, 2001). Os clones, cujo mRNA não são diferencialmente expressos entre as duas amostras de cDNAs marcados, terão uma cor intermediária entre o verde e o vermelho, aqui representados pela cor amarela. Os pontos onde não se observa nenhuma das três cores citadas correspondem aos genes que não possuem relação nenhuma com as situações estudadas.

Como dados numéricos, relacionados às medidas dos níveis de expressão dos genes são obtidos com variabilidade (da realização do experimento, ou do processo de obtenção das imagens, entre outros), métodos estatísticos são importantes para a análise dos dados, com o objetivo de identificar os possíveis genes que apresentam evidências para níveis de expressão diferentes devido apenas ao efeito de tratamento.

Os primeiros métodos estatísticos descritos para análise da expressão gênica foram discutidos por Schena *et al.*, (1995), Schena *et al.*, (1996) e DeRisi *et al.*, (1996).

Mais recentemente, diversos artigos abordam diferentes métodos estatísticos para análise da expressão gênica, entre eles, Baldi e Long (2001) com a utilização do teste t e uma abordagem bayesiana, Efron *et al.* (2001) através de uma abordagem bayesiana empírica, Do *et al.* (2002) com o modelo de misturas de processos Dirichlet, Medvedovic e Sivaganasan (2002) através de um modelo com mistura, baseado no modelo de misturas de processos Dirichlet.

Consideramos para análise da expressão gênica as situações controle (por exemplo, medidas de níveis de expressão das sondas de cDNA's provenientes de células sãs) e tratamentos (medidas de níveis de expressão das sondas de cDNA's em qualquer situação diferente do controle).

O objetivo de nossa pesquisa é comparar o desempenho e desenvolver métodos estatísticos, capazes de identificar com qualidade, genes que apresentam evidências para níveis

[4]RNA mensageiro, utilizado na produção das sondas.

de expressão diferentes, quando comparamos situações de interesse (tratamentos) com uma situação de controle.

Para as análises descritas nos capítulos seguintes, determinamos os logaritmos[5] das medidas dos níveis de expressão observadas para cada gene em cada situação, controle e tratamento. Assim, para cada gene, temos medidas de níveis de expressão, controle e tratamento, definidas no intervalo $(-\infty, \infty)$. Dessa forma, vamos supor que estas medidas transformadas foram geradas segundo uma distribuição normal. Esta suposição de normalidade é largamente utilizada na literatura, ver por exemplo Baldi e Long (2001), Efron *et al.* (2001), Do *et al.* (2002) e Medvedovic e Sivaganasan (2002).

Assim, para cada gene g, $g = 1, 2, ..., G$, temos um conjunto de variáveis observáveis $x^c_{g1}, ..., x^c_{gn_c}$ e $x^t_{g1}, ..., x^t_{gn_t}$, independentes, representando o logaritmo dos níveis de expressão do gene g na situação controle (c) e tratamento (t), respectivamente. Genes diferentes são tratados independentemente.

No Capítulo 2, descrevemos uma abordagem comumente utilizada para análise dos dados de expressão gênica e o teste t proposto por Baldi e Long (2001). Desenvolvemos para o teste t, um estudo de simulação para verificar seu comportamento na detecção de genes com evidências para níveis de expressão diferentes quando consideramos diferentes afastamentos na média, na variância ou em ambos, das observações de tratamento com relação as observações de controle. Aplicamos o teste t a dados de níveis de expressão reais, obtidos de um experimento realizado com as células da bactéria *Escherichia Coli* sob os padrões IHF^+ e IHF^- (ver Arfin *et al.*, 2000).

Baseados no trabalho de Baldi e Long (2001), propomos no Capítulo 3 um método de identificação de genes com e sem evidências para diferença, utilizando a distribuição de probabilidades das medidas de níveis de expressão observadas, para cada gene em cada situação. Pois assim, estaremos fazendo inferências tanto com relação à variação na média quanto com relação à variação na variância das medidas de tratamento com relação às medidas de controle. Para isto, consideramos os modelos, M_0 (medidas de níveis de expressão, tratamento e controle, sem evidência para diferença) e M_1 (medidas de níveis de expressão, tratamento e controle, com evidências para diferença). Para a seleção do modelo, M_0 ou M_1, que melhor explica as medidas de níveis de expressão observadas

[5] Aqui nos referimos aos logaritmos naturais.

para um determinado gene, utilizamos a inferência bayesiana paramétrica e os métodos de seleção de modelos: fator de Bayes e o critério DIC.

Desenvolvemos um estudo de simulação com objetivo de verificar o comportamento do fator de Bayes e do critério DIC na detecção de genes com evidências para níveis de expressão diferentes quando consideramos diferentes afastamentos na média e/ou variância das observações de tratamento com relação as observações de controle. Com o estudo de simulação podemos observar que o fator de Bayes e o critério DIC apresentam resultados semelhantes e detectam evidências para níveis de expressão diferentes, tanto com relação à variação na média quanto com relação à variação na variância, ou ambos.

Aplicamos o fator de Bayes e o critério DIC a dados reais, obtidos do experimento realizado com células da bactéria *Escherichia Coli*.

Na simulação e na aplicação para o fator de Bayes e para o critério DIC, utilizamos diferentes valores para os hiperparâmetros da distribuição *a priori* para a variância dos modelos M_0 e M_1, de cada um dos genes.

O fator de Bayes e o critério DIC, tanto na simulação quanto na aplicação, se mostraram sensíveis à escolha dos hiperparâmetros da distribuição *a priori* para as variância dos modelos M_0 e M_1. À medida que temos informação *a priori*, o fator de Bayes e o critério DIC apresentam um melhor desempenho na detecção dos genes com evidências para diferença. Mesmo com esta sensíbilidade, os resultados obtidos com o fator de Bayes e com o DIC se mstram melhores do que com os obtidos com o teste t.

Com o objetivo de melhorar as inferências sobre os possíveis genes que apresentam evidências para níveis de expressão diferentes e diminuir as restrições para as análises, no Capítulo 4, utilizamos a abordagem bayesiana semi-paramétrica, conhecida como modelo de misturas de processo Dirichlet (modelo MPD). Primeiramente, revisamos a definição e algumas propriedades da distribuição Dirichlet. Baseados em Ferguson (1973), definimos o processo Dirichlet, e algumas de suas propriedades, tais como, média, variância e distribuição *a posteriori*.

Para Ibrahim (2001) o processo Dirichlet é o mais popular processo *a priori* em inferência bayesiana não paramétrica.

Com a utilização do processo Dirichlet *a priori*, atribuímos incerteza à uma distribuição de probabilidades e com isso obtemos um modelo mais flexível e abrangente.

Pois, em vez de afirmar que os efeitos aleatórios, da variável de interesse, pertencem apenas ao conjunto de distribuições normais, o modelo permite que estes efeitos se adaptem à distribuições desconhecidas, que podem ser assimétricas, multímodais ou simplesmente ter a forma paramétrica de uma distribuição normal.

Descrevemos o modelo de misturas de processos Dirichlet, que pode ser utilizado com sucesso na análise de dados com mistura de distribuições e um algoritmo, baseado no método *Gibbs Sampling* (ver Gelfand e Smith, 1990), para se obter amostras das distribuições condicionais de um modelo de misturas de processos Dirichlet conjugado. Realizamos um estudo de simulação no qual aplicamos o modelo MPD na estimação da média de uma distribuição normal com variância conhecida e na estimação da média e variância de uma distribuição normal, quando estes dois parâmetros são desconhecidos.

O algoritmo *Gibbs Sampling* descrito para o modelo de misturas de processos Dirichlet conjugado é usado por Escobar (1994) e por Escobar e West (1995). Segundo Neal (1998), este algoritmo produz uma cadeia de Markov ergódica[6], porém a obtenção de convergência pode ser lenta.

Aplicamos o modelo MPD e o método *Gibbs Sampling* na análise da expressão gênica. Para as inferências sobre os genes que apresentam evidências para níveis de expressão diferentes, consideramos a proporção de vezes que as médias $\bar{x}_{gc}$ e $\bar{x}_{gt}$, das observações de controle e tratamento de um determinado gene g, respectivamente, foram consideradas como sendo geradas de uma mesma distribuição normal pelo modelo MPD, durante as p seqüências geradas pelo método *Gibbs Sampling*, para $g = 1, 2, ..., G$.

Desenvolvemos um estudo de simulação para verificarmos o comportamento do modelo MPD na detecção de genes com evidências para diferença, quando consideramos diferentes afastamentos na média e na variância (ou ambos) das observações de tratamento com relação as observações de controle.

Aplicamos o modelo MPD aos dados obtidos do experimento realizado com células da bactéria *Escherichia Coli*.

Para simulação e aplicação, consideramos diferentes valores para os hiperparâmetros das distribuições *a priori*.

Tanto na simulação quanto na aplicação, o modelo MPD se mostrou sensível a escolha

[6] Irredutível, aperiódica e recorrente positiva.

dos hiperparâmetros da distribuição *a priori* para a variância, levando a diferentes conclusões, ou seja, a escolha de diferentes genes g com evidências para níveis de expressão diferentes, $g = 1, 2, ..., G$. Isto nos mostra que devemos ter uma certa atenção e alguma informação para determinarmos estes hiperparâmetros e obtermos resultados satisfatórios.

À medida que temos informação *a priori* para a variância, o modelo MPD se mostra com um melhor desempenho na detecção dos genes com evidência para diferença.

Uma abordagem também utilizada considera a identificação e análise de grupos (ou *clusters*) de genes com níveis de expressão similares devido ao efeito de tratamento ao invés da análise de um individual gene de cada vez (ver Medvedovic e Sivaganesan, 2002).

Dessa forma, no Capítulo 5 propomos um modelo bayesiano, baseado em um modelo com mistura infinita que é equivalente ao modelo MPD e descrevemos um algoritmo, baseado no método *Gibbs Sampling*, para se obter amostras das distribuições condicionais de um modelo com mistura infinita. Aplicamos este modelo na análise da expressão gênica com o objetivo de identificar grupos (ou *clusters*) de genes com níveis de expressão similares.

Identificados os grupos (ou *clusters*) de genes com níveis de expressão similares, consideramos para cada grupo os modelos, M_0 e M_1, M_0 se o grupo é composto por medidas sem evidências para diferença e M_1 se o grupo é composto por medidas com evidências para diferença. Para detectar se o grupo é composto por medidas de níveis de expressão que apresentam ou não evidências para diferença, utilizamos a abordagem bayesiana e o critério DIC.

Desenvolvemos um estudo de simulação para verificarmos o comportamento do modelo com mistura infinita na identificação de grupos de genes com níveis de expressão similares e do critério DIC na identificação dos grupos compostos por medidas de níveis de expressão com evidências para diferença, quando consideramos diferentes afastamentos na média e na variância (ou ambos) das observações de tratamento com relação as observações de controle.

Aplicamos a modelagem com misturas infinita e o critério DIC aos dados obtidos do experimento realizado com células da bactéria *Escherichia Coli*.

Para a simulação e para a aplicação, consideramos para formação dos grupos de genes com níveis de expressão similares, diferentes valores para os hiperparâmetros da

distribuição *a priori* para a variância. Para aplicação do critério DIC utilizamos os mesmos hiperparâmetros utilizados na formação dos grupos, pois o critério DIC não se mostrou sensível aos hiperparâmetros. Isto é, o critério DIC independentemente do valor escolhido para os hiperparâmetros da distribuição *a priori* para variância, detecta sempre os mesmos grupos com evidências para diferença.

Tanto na simulação quanto na aplicação o modelo com mistura infinita se mostrou sensível à escolha dos hiperparâmetros da distribuição *a priori* para a variância na a formação dos grupos (ou *clusters*). Mudando o valor dos hiperparâmetros os genes com níveis de expressão na fronteira dos grupos formados trocam de grupos. Logo, devemos ter certa atenção e alguma informação para determinarmos os valores dos hiperparâmetros e obtermos resultados satisfatórios. Se possível devemos nos basear na opinião de um especialista ou utilizar métodos empíricos para determinar os valores dos hiperparâmetros.

No Capítulo 6, fazemos algumas considerações finais sobre os métodos estatísticos aplicados na análise da expressão gênica, apresentados nesta dissertação.

Capítulo 2

Abordagem usual e Teste t

Uma abordagem comumente utilizada para análise da expressão gênica, considera que um gene g, apresenta evidências para níveis de expressão diferentes se o logaritmo da razão entre a média das observações de tratamento e a média das observações de controle, em valor absoluto, variar mais do que um valor de referência, tipicamente 2. Isto é, se o valor absoluto obtido para o logaritmo da razão entre a média das observações de tratamento e a média das observações de controle, for maior que o valor de referência, o gene g é considerado com evidências para diferença, para $g = 1, 2, ..., G$ (ver, Felix *et al.*, 2001, Baldi e Long, 2001 e Do *et al.*, 2002).

Uma outra abordagem também muito utilizada, considera para um gene g, que o logaritmo das observações de controle foram geradas de uma distribuição normal com média μ_{gc} e variância σ^2_{gc},

$$x^c_{g1}, ..., x^c_{gn_c} \sim N\left(\mu_{gc}, \sigma^2_{gc}\right)$$

e o logaritmo das observações de tratamento foram geradas de uma distribuição normal com média μ_{gt} e variância σ^2_{gt},

$$x^t_{g1}, ..., x^t_{gn_t} \sim N\left(\mu_{gt}, \sigma^2_{gt}\right),$$

onde n_c e n_t correspondem às quantidades de observações de níveis de expressão na situação de controle e na situação de tratamento, respectivamente, para $g = 1, 2, ..., G$ (ver Baldi e Long, 2001).

Para determinar se o gene g apresenta ou não evidências para níveis de expressão

diferentes, utiliza-se o teste de hipóteses sob a forma

$$H_0 : \mu_{gc} = \mu_{gt} \; versus \; H_1 : \mu_{gc} \neq \mu_{gt} \tag{2.1}$$

para g, $g = 1, 2, ..., G$.

Baldi e Long (2001) utilizam o teste t, para cada gene g, baseados no teste de hipóteses como em (2.1), e as médias e variâncias amostrais são utilizadas para determinar se o gene g apresenta evidências para diferença, sob a forma

$$t_g = \frac{\bar{x}_{gc} - \bar{x}_{gt}}{\sqrt{\frac{s^2_{gc}}{n_c} + \frac{s^2_{gt}}{n_t}}}$$

com t_g seguindo uma distribuição t-Student, com p graus de liberdade,

$$p = \frac{\left[\frac{s^2_{gc}}{n_c} + \frac{s^2_{gt}}{n_t}\right]^2}{\frac{\left(\frac{s^2_{gc}}{n_c}\right)^2}{n_c - 1} + \frac{\left(\frac{s^2_{gt}}{n_t}\right)^2}{n_t - 1}}, \tag{2.2}$$

onde $\bar{x}_{gc}$ é a média amostral das observações de controle para o gene g, dada por

$$\bar{x}_{gc} = \frac{1}{n_c} \sum_{i=1}^{n_c} x^c_{gi},$$

$\bar{x}_{gt}$ é a média amostral das observações de tratamento, dada por

$$\bar{x}_{gt} = \frac{1}{n_t} \sum_{i=1}^{n_t} x^t_{gi},$$

s^2_{gc} é a variância amostral das observações de controle, dada por

$$s^2_{gc} = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} \left(x^c_{gi} - \bar{x}_{gc}\right)^2$$

e s^2_{gt} é a variância amostral das observações de tratamento, dada por

$$s^2_{gt} = \frac{1}{n_t - 1} \sum_{i=1}^{n_t} \left(x^t_{gi} - \bar{x}_{gt}\right)^2.$$

Fixado um nível de significância α, se $\mid t_g \mid$ é maior que o valor de referência $t_{1-\frac{\alpha}{2},p}$ $\left(\text{quantil } 1-\frac{\alpha}{2} \text{ da distribuição } t-Student \text{ com } p \text{ graus de liberdade}\right)$ então há evidências de que o gene g apresenta níveis de expressão significativamente diferentes, quando comparamos a situação de tratamento com a situação controle, para $g = 1, 2, ..., G$.

Um problema que surge para a aplicação do teste t aos dados de expressão gênica, é o tamanho das amostras n_c e n_t, para um gene g, que geralmente são pequenas.

É importante notar que buscamos de fato uma ferramenta estatística que separe as distribuições das observações de níveis de expressão relativas a controle e tratamento, quando estas são diferentes. Como veremos a seguir o teste t separa eficientemente estas distribuições quando observamos diferença de médias, entre tratamento e controle, mas com variâncias, tratamento e controle, razoavelmente estáveis. Quando variações na média de tratamento com relação a média de controle está acompanhada de aumento na variância de tratamento com relação a variância de controle o teste t não funciona adequadamente, embora tenha havido mudança na distribuição das observações de expressão gênica de tratamento com relação a de controle.

2.1 Simulação

Desenvolvemos um estudo de simulação, para verificarmos o comportamento do teste t, como proposto por Baldi e Long (2001), na detecção de genes com evidências para níveis de expressão diferentes, quando consideramos diferentes afastamentos na média e na variância (ou ambos) das observações de tratamento com relação às observações de controle.

Para gerar os dados de expressão gênica consideramos uma amostra controle com o logaritmo dos níveis expressão provenientes de uma distribuição normal, $N(\mu_c, \sigma_c^2)$, e uma amostra tratamento, com o logaritmo dos níveis de expressão provenientes de uma distribuição normal, $N(\mu_t, \sigma_t^2)$, onde $\mu_t = \mu_c + \delta$ e $\sigma_t^2 = (\gamma\sigma_c)^2$.

O passo seguinte é considerar variações em δ e γ. Dessa forma, a partir de $\delta = 0$, aumentando ou diminuindo o valor de δ e aumentando ou diminuindo o valor de γ esperamos detectar uma quantidade maior de genes (simulados) com evidências para níveis de expressão diferentes. Pois estaremos afastando a distribuição das observações

de tratamento da distribuição das observações de controle.

Para cada variação de δ e γ considerados, aplicamos o teste t, com um nível de significância $\alpha = 0,05$, para detectar os genes (simulados) que apresentam evidências para níveis de expressão diferentes.

Consideramos para a simulação os seguintes passos:

1 - Simulamos 1000 genes g, $g = 1, 2, ..., 1000$,

2 - Geramos para cada um dos 1000 genes, 5 observações de controle, com distribuição $N(\mu_c, \sigma_c^2)$ e 5 observações de tratamento, com distribuição $N(\mu_t, \sigma_t^2)$.

3 - Fixamos como valor para $\mu_c = -0,036$ e $\sigma_c^2 = 0,04$, baseados nas observações do experimento realizado com as células da bactéria *Escherichia Coli* (ver Arfin *et al.*, 2000).

A Tabela 1 mostra as quantidades de genes (simulados) detectados com evidências para níveis de expressão significativamente diferentes, para cada variação de δ e γ (ou ambos) considerados. Cada linha mostra as quantidades detectadas, quando consideramos γ fixo e variamos δ. Cada coluna mostra as quantidades detectadas, quando consideramos δ fixo e variamos γ.

Para γ fixo (cada linha) a medida que afastamos a média da distribuição das observações de tratamento da média da distribuição das observações de controle (a partir de $\delta = 0$, aumentando ou diminuindo o valor de δ), o teste t detecta uma quantidade maior de genes com evidências para níveis de expressão diferentes do que a variação δ anterior. Isto nos mostra que o teste t é sensível à variação na média, quando γ é fixo, detectando os genes com evidência para níveis de expressão diferentes.

Já para δ fixo (cada coluna), a partir de $\gamma = 0,25$, igualando ou aumentando a variância da distribuição das observações de tratamento com relação a distribuição das observações de controle (aumentamos o valor de γ), o teste t detecta menos genes com evidências para níveis de expressão diferentes. Por exemplo, para $\delta = 0,5$, $\gamma = 1$ e $\gamma = 9$, respectivamente, 916 e 207 genes foram detectados com evidências para diferença. Ocorre o contrário do que esperávamos, que seria detectar uma quantidade maior ou igual de genes com evidências para diferença à medida que aumentamos o valor de γ.

Dessa forma, notamos que a utilização do teste t, proposto por Baldi e Long (2001), para a análise da expressão gênica se mostra uma ferramenta estatística inadequada para detectar alterações na média quando esta está acompanhada de um aumento na variância

das observações de tratamento com relação a variância das observações de controle. Para variâncias estáveis ($\gamma = 0,25$ ou $\gamma = 1$) os resultados obtidos são adequados, pois para $\delta \neq 0$ a quantidade detectada com evidências para diferença e próxima ($\delta = \pm 0,5$) ou igual ($\delta = \pm 0$ ou $\delta = 1$) a 1000 genes (simulados).

Tabela 1: Quantidades detectadas com evidências para diferença pelo teste t.

	δ						
γ	-1,0	-0,8	-0,5	0	0,5	0,8	1,0
0,25	1000	1000	982	55	976	1000	1000
1	1000	1000	930	46	916	997	1000
4	977	895	557	45	563	910	976
9	797	630	291	43	302	636	822
16	585	426	196	41	207	412	586

2.2 Aplicação

Aplicamos o teste t na análise da expressão gênica, utilizando as medidas de níveis de expressão obtidos do experimento realizado com células da bactéria *Escherichia Coli,* com relação aos padrões IHF^+ e IHF^- (ver Arfin *et al.*, 2000).

Para a análise dispomos de 434 genes, com cinco medidas de níveis de expressão para a situação de tratamento (t) e cinco medidas de níveis de expressão para a situação de controle (c).

Para aplicação do teste t, consideramos os níveis de significância $\alpha = 0,10$ e $\alpha = 0,05$.

As Figuras 2 e 3, mostram as médias e variâncias controle e tratamento observadas, destacando os genes detectados com evidências para níveis de expressão diferentes. Nestas Figuras os pontos • indicam os genes que não foram detectados com evidências para diferença e os sinais + indicam os genes que foram detectados com evidências para diferença. Note que os genes com média de tratamento e de controle distantes da reta $y = x$ não foram detectados.

Com $\alpha = 0,10$ (Figura 2), 51 genes foram detectados com evidências para diferença, enquanto que com $\alpha = 0,05$ (Figura 3) 21 genes foram detectados com evidências para diferença.

Como notado na simulação, o teste t detecta evidências para diferença somente quando temos uma diferença na média da distribuição das observações de tratamento com relação a média da distribuição das observações de controle com as variâncias destas duas distribuições sendo razoavelmente estáveis. Isto explica, porque os genes com média de tratamento e de controle distantes da reta $y = x$ não foram detectados e genes com média de tratamento e de controle mais próximos a reta $y = x$ foram detectados. Por exemplo, o gene 277 apresenta uma diferença tanto na média quanto na variância e não foi detectado com evidências para diferença e o gene 227 que apresenta média de tratamento e de controle mais próximos à reta $y = x$ foi detectado com evidências para diferença.

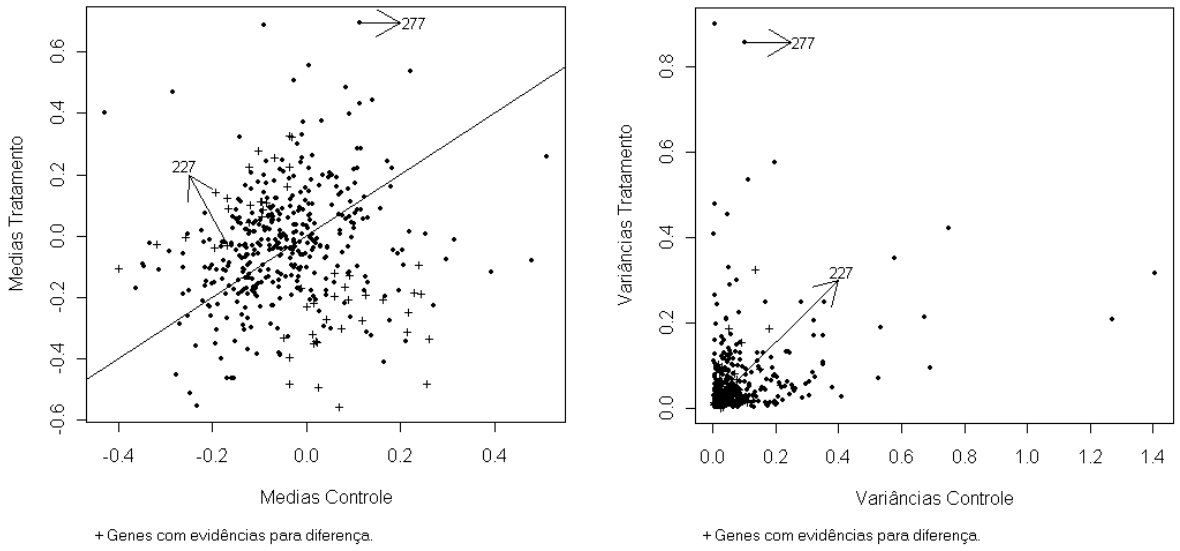

Figura 2: Médias e variâncias controle e tratamento, com $\alpha = 0,10$.

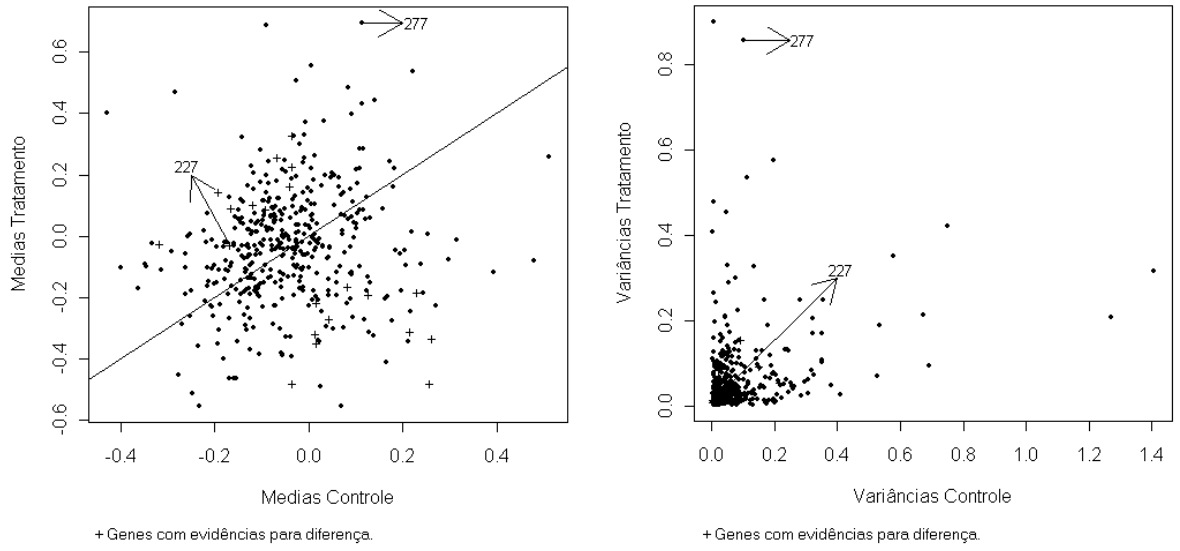

Figura 3: Médias e variâncias controle e tratamento, com $\alpha = 0,05$.

A Tabela 2 mostra o número e as médias e variâncias amostrais controle e tratamento, dos 51 genes detectados com evidências para níveis de expressão diferentes pelo teste t. Os genes indicados com ∘ sao os detectados quando utilizamos o nível de significância $\alpha = 0,05$. Observe que todos os genes foram detectados com evidências devido a uma diferença presente nas médias, controle e tratamento, pois as variâncias são próximas.

Tabela 2: Genes identificados com evidência para diferença pelo teste t.

N^o gene	Média cont.	Variância cont.	Média trat.	Variância trat.
∘02	-0,0412	0,0069	0,1643	0,0141
∘12	-0,0348	0,0145	0,2283	0,0231
17	-0,0363	0,0401	-0,3933	0,1173
34	-0,1224	0,0205	0,0467	0,0116
59	-0,3980	0,0650	-0,1013	0,0248
∘84	-0,0359	0,0792	-0,4785	0,0473
∘86	-0,0674	0,0350	0,2573	0,0423
88	-0,1023	0,0158	0,2827	0,1035
∘98	-0,1670	0,0323	0,0937	0,0269
114	0,0907	0,0348	-0,1232	0,0179
128	0,0202	0,0930	-0,3389	0,0241
∘134	0,2603	0,2152	-0,3331	0,0567
169	-0,0487	0,0414	-0,3261	0,0527
∘176	-0,0357	0,0298	0,3309	0,0402
179	0,0705	0,1365	-0,5547	0,3276
∘183	0,0439	0,0210	-0,2685	0,0350
202	0,0891	0,0729	-0,2033	0,0126
∘212	0,0158	0,0222	-0,3466	0,0865
∘227	-0,1684	0,0039	-0,0265	0,0102
233	0,0249	0,1770	-0,4876	0,1894
238	0,0594	0,0234	-0,1907	0,0538
∘244	0,1267	0,0166	-0,1890	0,0260

Nº gene	Média cont.	Variância cont.	Média trat.	Variância trat.
°248	0,2133	0,1489	-0,3095	0,0940
°254	-0,1924	0,0293	0,1445	0,0255
°259	0,0822	0,0139	-0,1641	0,0283
269	-0,0108	0,0250	-0,1651	0,0060
271	-0,1217	0,0330	0,2292	0,0919
283	-0,0959	0,0138	0,0685	0,0136
286	0,0008	0,0434	-0,2262	0,0178
288	0,0607	0,0315	-0,1172	0,0040
311	0,1195	0,1080	-0,2700	0,0167
318	-0,0312	0,0299	0,3249	0,0988
°345	0,0161	0,0113	-0,2167	0,0112
352	0,2392	0,1095	-0,0896	0,0219
359	0,0224	0,0671	-0,3436	0,0824
360	0,2449	0,1277	-0,1841	0,0318
369	0,0744	0,0843	-0,2979	0,0440
°375	0,0133	0,0085	-0,3156	0,0238
376	0,2173	0,0489	-0,2458	0,1899
°383	0,2566	0,0903	-0,4787	0,1561
°386	-0,0912	0,0113	0,0906	0,0173
389	-0,1943	0,0239	-0,0353	0,0022
°393	-0,3188	0,0399	-0,0242	0,0374
409	-0,0844	0,0225	0,1123	0,0285
°410	-0,1181	0,0046	0,1035	0,0161
412	-0,0947	0,0136	0,1154	0,0282
416	0,1632	0,0758	-0,2045	0,0708
423	-0,2562	0,0376	-0,0023	0,0397
427	0,2287	0,0586	-0,1830	0,0409
431	-0,0963	0,0198	0,1138	0,0231
432	-0,1694	0,0428	0,1265	0,0598

Capítulo 3

Abordagem bayesiana paramétrica, Fator de Bayes e DIC

Neste capítulo, consideramos para a análise da expressão gênica os modelos M_0 e M_1. No modelo M_0 os níveis de expressão são provenientes de uma mesma distribuição de probabilidades e no modelo M_1 os níveis de expressão são provenientes de distribuições de probabilidades diferentes. Para selecionar o modelo que melhor explica os níveis de expressão observados e conseqüentemente, determinar se um gene apresenta ou não evidências para níveis de expressão diferentes, utilizamos a inferência bayesiana paramétrica e os métodos de seleção de modelos, fator de Bayes e critério DIC.

Nesta abordagem, em ambos os modelos, genes diferentes são considerados independentes.

3.1 Modelo para a expressão gênica

No modelo M_0, para um determinado gene g, consideramos que o logaritmo das medidas dos níveis de expressão observadas na situação de controle e de tratamento foram geradas de uma mesma distribuição de probabilidades,

$$x^c_{g1}, ..., x^c_{gn_c}, x^t_{g1}, ..., x^t_{gn_t} = x_{g1}, ..., x_{gn_c}, x_{gn_c+1}..., x_n \sim N\left(\mu_g, \sigma^2_g\right),$$

onde $x_{gn_c+1} = x^t_{g1}, ..., x_n = x^t_{gn_t}$, com $n = n_c + n_t$, para $g = 1, 2, ..., G$.

Para este modelo, denotamos por

$$D_g = \left\{x^c_{g1}, ..., x^c_{gn_c}, x^t_{g1}, ..., x^t_{gn_t}\right\} = \{x_{g1}, ..., x_{gn_c}, x_{gn_c+1}..., x_n\}$$

o conjunto das medidas dos níveis de expressão observados na situação de controle e de tratamento para o gene g, e a quantidade de medidas que compõem D_g é $n = n_c + n_t$.

No modelo M_1, para um determinado gene g, consideramos que o logaritmo das medidas dos níveis de expressão observadas na situação de controle e de tratamento foram geradas de distribuições de probabilidades diferentes,

$$x^c_{g1}, ..., x^c_{gn_c} \sim N\left(\mu_{gc}, \sigma^2_{gc}\right) \text{ e } x^t_{g1}, ..., x^t_{gn_t} \sim N\left(\mu_{gt}, \sigma^2_{gt}\right).$$

Para este modelo, denotamos por

- $D_{gc} = \left\{x^c_{g1}, ..., x^c_{gn_c}\right\}$: o conjunto das medidas dos níveis de expressão observadas para o gene g na situação de controle,

- $D_{gt} = \left\{x^t_{g1}, ..., x^t_{gn_t}\right\}$: o conjunto das medidas dos níveis de expressão observadas para o gene g na situação de tratamento, e

- n_c e n_t: correspondem as quantidades de medidas de níveis de expressão observadas que compõem, D_{gc} e D_{gt}, respectivamente.

Para cada modelo, M_0 e M_1, temos as verossimilhanças, dadas por

$$L_{M_0}\left(\mu_g, \sigma^2_g | D_g\right) = \prod_{i=1}^{n} f_{M_0}(\mu_g, \sigma^2_g) = \left(\frac{1}{\sqrt{2\pi\sigma^2_g}}\right)^n \exp\left\{-\frac{1}{2\sigma^2_g}\sum_{i=1}^{n}\left(x_{gi} - \mu_g\right)^2\right\}$$

que pode ser escrita sob a forma,

$$L_{M_0}\left(\mu_g, \sigma^2_g | D_g\right) \propto \left(\sigma^2_g\right)^{-\frac{n}{2}} \exp\left\{-\frac{n(\bar{x}_g - \mu_g)^2 + (n-1)s^2_g}{2\sigma^2_g}\right\} \tag{3.1}$$

e

$$L_{M_1}\left(\mu_{gc}, \sigma^2_{gc}, \mu_{gt}, \sigma^2_{gt} | D_{gc}, D_{gt}\right) = L_c\left(\mu_{gc}, \sigma^2_{gc} | D_{gc}\right) L_t\left(\mu_{gt}, \sigma^2_{gt} | D_{gt}\right) \tag{3.2}$$

onde,

$$L_c\left(\mu_{gc}, \sigma^2_{gc} | D_{gc}\right) \propto \left(\sigma^2_{gc}\right)^{-\frac{n_c}{2}} \exp\left\{-\frac{n_c(\bar{x}_{gc} - \mu_{gc})^2 + (n_c - 1)s^2_{gc}}{2\sigma^2_{gc}}\right\}$$

e

$$L_t\left(\mu_{gt}, \sigma^2_{gt}|D_{gt}\right) \propto \left(\sigma^2_{gt}\right)^{-\frac{n_t}{2}} \exp\left\{-\frac{n_t(\bar{x}_{gt}-\mu_{gt})^2 + (n_t-1)s^2_{gt}}{2\sigma^2_{gt}}\right\}$$

para $g = 1, 2, ..., G$.

Considerar se há ou não evidências para níveis de expressão diferentes para um determinado gene g, equivale a considerar o modelo M_0 ou M_1 como sendo o modelo que melhor explica às medidas dos níveis de expressão observados. Dessa forma consideramos a abordagem bayesiana paramétrica e os métodos de seleção de modelos, fator de Bayes e critério DIC, para selecionar o modelo que melhor explica os níveis de expressão observados e assim, determinar se o gene g apresenta ou não evidências para níveis de expressão diferentes entre a situação de controle e de tratamento, para $g = 1, 2, ..., G$.

3.2 Abordagem Bayesiana

Para a abordagem Bayesiana, devemos especificar para as situações dos modelos M_0 e M_1, distribuições *a priori* para seus respectivos parâmetros μ_g, σ^2_g e μ_{gc}, σ^2_{gc}, μ_{gt}, σ^2_{gt}, para $g = 1, 2, ..., G$ e fazemos inferências baseados em suas distribuições *a posteriori*.

Aqui supomos independência *a priori* entre os parâmetros dos modelos M_0 e M_1 e também entre os parâmetros associados a situação de controle e tratamento do modelo M_1. Como resultado desta última suposição temos que a distribuição *a posteriori* para a situação do modelo M_1 se fatora para os dados relativos a controle e tratamento.

Para o modelo M_0, consideramos o conjunto de hiperparâmetros $\varpi_0 = (\mu_0, \lambda, \alpha, \beta)$ e as distribuições *a priori* usuais, normal e gama inversa dadas por

$$\pi(\mu_g|\sigma^2_g) = N\left(\mu_0, \frac{\sigma^2_g}{\lambda}\right)$$

ou

$$\pi(\mu_g|\sigma^2_g) \propto (\sigma^2_g)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda}{2\sigma^2_g}\left(\mu_g - \mu_0\right)^2\right\}$$

e

$$\pi(\sigma^2_g) = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

ou

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

A distribuição *a priori* conjunta, é dada por

$$\pi(\mu_g, \sigma_g^2) = \pi(\mu_g|\sigma_g^2)\pi(\sigma_g^2),$$

ou

$$\pi(\mu_g, \sigma_g^2) \propto \left(\sigma_g^2\right)^{-\left(\frac{\alpha+1}{2}+1\right)} \exp\left\{-\frac{\lambda}{2\sigma_g^2}\left(\mu_g - \mu_0\right)^2 - \frac{\beta}{2\sigma_g^2}\right\}. \tag{3.3}$$

Para o modelo M_1, consideramos o conjunto de hiperparâmetros, $\varpi_1 = (\mu_{0c}, \lambda_c, \alpha_c, \beta_c, \mu_{0t}, \lambda_t, \alpha_t, \beta_t)$ e as distribuições *a priori* usuais,

$$\begin{aligned}
\pi(\mu_{gc}|\sigma_{gc}^2) &= N\left(\mu_{0c}, \frac{\sigma_{gc}^2}{\lambda_c}\right) \Longrightarrow \pi(\mu_{gc}|\sigma_{gc}^2) \propto (\sigma_{gc}^2)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda_c}{2\sigma_{gc}^2}\left(\mu_{gc} - \mu_{0c}\right)^2\right\} \\
\pi(\mu_{gt}|\sigma_{gt}^2) &= N\left(\mu_{0t}, \frac{\sigma_{gt}^2}{\lambda_t}\right) \Longrightarrow \pi(\mu_{gt}|\sigma_{gt}^2) \propto (\sigma_{gt}^2)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda_t}{2\sigma_{gt}^2}\left(\mu_{gt} - \mu_{0t}\right)^2\right\}
\end{aligned}$$

e

$$\begin{aligned}
\pi(\sigma_{gc}^2) &= IG\left(\frac{\alpha_c}{2}, \frac{\beta_c}{2}\right) \Longrightarrow \pi(\sigma_{gc}^2) \propto \left(\sigma_{gc}^2\right)^{-\left(\frac{\alpha_c}{2}+1\right)} \exp\left\{-\frac{\beta_c}{2\sigma_{gc}^2}\right\} \\
\pi(\sigma_{gt}^2) &= IG\left(\frac{\alpha_t}{2}, \frac{\beta_t}{2}\right) \Longrightarrow \pi(\sigma_{gt}^2) \propto \left(\sigma_{gt}^2\right)^{-\left(\frac{\alpha_t}{2}+1\right)} \exp\left\{-\frac{\beta_t}{2\sigma_{gt}^2}\right\}.
\end{aligned}$$

Com distribuição *a priori* conjunta, dada por

$$\pi(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2) = \pi(\mu_{gc}|\sigma_{gc}^2)\pi(\sigma_{gc}^2)\pi(\mu_{gt}|\sigma_{gt}^2)\pi(\sigma_{gt}^2),$$

ou

$$\begin{aligned}
\pi(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2) \propto\ & \left(\sigma_{gc}^2\right)^{-\left(\frac{\alpha_c+1}{2}+1\right)} \left(\sigma_{gt}^2\right)^{-\left(\frac{\alpha_t+1}{2}+1\right)} \\
& \exp\left\{-\frac{\lambda_c}{2\sigma_{gc}^2}\left(\mu_{gc} - \mu_{0c}\right)^2 - \frac{\lambda_t}{2\sigma_{gt}^2}\left(\mu_{gt} - \mu_{0t}\right)^2 - \frac{\beta_c}{2\sigma_{gc}^2} - \frac{\beta_t}{2\sigma_{gt}^2}\right\}.
\end{aligned} \tag{3.4}$$

Os modelos M_0 e M_1 apresentados acima seguem a proposta de Baldi e Long (2001).

Aplicando o teorema de Bayes, temos que a distribuição *a posteriori* para o modelo

M_0 é dada por

$$
\begin{aligned}
\pi\left(\mu_g, \sigma_g^2 | D_g, \varpi_0\right) &\propto L_{M_0}\left(\mu_g, \sigma_g^2 | D_g\right) \pi(\mu_g|\sigma_g^2)\pi\left(\sigma_g^2\right) \qquad (3.5)\\
&\propto \left(\sigma_g^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{n(\bar{x}_g-\mu_g)^2+(n-1)s_g^2}{2\sigma_g^2}\right\}\\
&\quad \left(\sigma_g^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda}{2\sigma_g^2}\left(\mu_g-\mu_0\right)^2\right\}\left(\sigma_g^2\right)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_g^2}\right\}\\
&\propto \exp\left\{-\frac{n(\bar{x}_g-\mu_g)^2+(n-1)s_g^2}{2\sigma_g^2}-\frac{\lambda}{2\sigma_g^2}\left(\mu_g-\mu_0\right)^2\right\}\\
&\quad \left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_g^2}\right\}\\
&\propto \exp\left\{-\frac{n(\bar{x}_g-\mu_g)^2}{2\sigma_g^2}-\frac{\lambda}{2\sigma_g^2}\left(\mu_g-\mu_0\right)^2\right\}\\
&\quad \left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_g^2}-\frac{(n-1)s_g^2}{2\sigma_g^2}\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma_g^2}\left[n(\bar{x}_g^2-2\mu_g\bar{x}_g+\mu_g^2)+\lambda\left(\mu_g^2-2\mu_g\mu_0+\mu_0^2\right)\right]\right\}\\
&\quad \left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma_g^2}\left(\beta+(n-1)s_g^2\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma_g^2}\left(n\bar{x}_g^2-2\mu_g n\bar{x}_g+n\mu_g^2+\lambda\mu_g^2-2\mu_g\lambda\mu_0+\lambda\mu_0^2\right)\right\}\\
&\quad \left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma_g^2}\left(\beta+(n-1)s_g^2\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2\sigma_g^2}\left(\mu_g^2(\lambda+n)-2\mu_g(n\bar{x}_g+\lambda\mu_0)\right)\right\}\\
&\quad \left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma_g^2}\left(\beta+(n-1)s_g^2+n\bar{x}_g^2+\lambda\mu_0^2\right)\right\}\\
&\propto \exp\left\{-\frac{\lambda+n}{2\sigma_g^2}\left(\mu_g^2-2\mu_g\left(\frac{n\bar{x}_g+\lambda\mu_0}{\lambda+n}\right)\right)\right\}\\
&\quad \left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma_g^2}\left(\beta+(n-1)s_g^2+n\bar{x}_g^2+\lambda\mu_0^2\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2\frac{\sigma_g^2}{\lambda+n}}\left(\mu_g-\frac{n\bar{x}_g+\lambda\mu_0}{\lambda+n}\right)^2\right\}\left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)}\\
&\quad \exp\left\{-\frac{1}{2\sigma_g^2}\left(\beta+(n-1)s_g^2+n\bar{x}_g^2+\lambda\mu_0^2-\frac{(n\bar{x}_g+\lambda\mu_0)^2}{\lambda+n}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2\frac{\sigma_g^2}{\lambda+n}}\left(\mu_g-\frac{n\bar{x}_g+\lambda\mu_0}{\lambda+n}\right)^2\right\}\left(\sigma_g^2\right)^{-\left(\frac{\alpha+n+1}{2}+1\right)}
\end{aligned}
$$

$$\exp\left\{-\frac{1}{\sigma_g^2}\left(\frac{\beta\left(\lambda+n\right)+\left(\lambda+n\right)\left(n-1\right)s_g^2+n\lambda\left(\bar{x}_g-\mu_0\right)^2}{2\left(\lambda+n\right)}\right)\right\}$$
$$= N\left(\mu_n,\frac{\sigma_g^2}{\lambda_n}\right)IG\left(\frac{\alpha_n}{2},\frac{\beta_n}{2}\right)$$

onde

$$\begin{aligned}\mu_n &= \frac{n}{\lambda\ +\ n}\bar{x}_g+\frac{\lambda}{\lambda\ +\ n}\mu_0\\ \lambda_n &= \lambda+n\\ \alpha_n &= \alpha+n+1\\ \beta_n &= \frac{\beta\lambda_n+\lambda_n(n-1)s_g^2+n\lambda(\bar{x}_g-\mu_0)^2}{\lambda_n}.\end{aligned}$$

A distribuição *a posteriori* para o modelo M_1 é dada por

$$\begin{aligned}\pi\left(\mu_{gc},\sigma_{gc}^2,\mu_{gt},\sigma_{gt}^2|D_{gc},D_{gt},\varpi_1\right) &= N\left(\mu_{cn},\frac{\sigma_{gc}^2}{\lambda_{cn}}\right)IG\left(\frac{\alpha_{cn}}{2},\frac{\beta_{cn}}{2}\right)\\ &\quad N\left(\mu_{tn},\frac{\sigma_{gt}^2}{\lambda_{tn}}\right)IG\left(\frac{\alpha_{tn}}{2},\frac{\beta_{tn}}{2}\right)\end{aligned} \tag{3.6}$$

onde

$$\begin{aligned}\mu_{cn} &= \frac{n_c}{\lambda_{_c}\ +\ n_c}\bar{x}_{gc}+\frac{\lambda_{_c}}{\lambda_c\ +\ n_c}\mu_{0c}\\ \lambda_{cn} &= \lambda_{_c}+n_c\\ \alpha_{cn} &= \alpha_c+n_c+1\\ \beta_{cn} &= \frac{\beta_c\lambda_{cn}+\lambda_{cn}(n_c-1)s_{gc}^2+n_c\lambda_c(\bar{x}_{gc}-\mu_{0c})^2}{\lambda_{cn}}\end{aligned}$$

e

$$\begin{aligned}\mu_{tn} &= \frac{n_t}{\lambda_t\ +\ n_t}\bar{x}_{gt}+\frac{\lambda_t}{\lambda_t\ +\ n_t}\mu_{0t}\\ \lambda_{tn} &= \lambda_t+n_t\\ \alpha_{tn} &= \alpha_t+n_t+1\\ \beta_{tn} &= \frac{\beta_t\lambda_{tn}+\lambda_{tn}(n_t-1)s_{gt}^2+n_t\lambda_t(\bar{x}_{gt}-\mu_{0t})^2}{\lambda_{tn}}.\end{aligned}$$

3.3 Seleção de Modelos: Fator de Bayes

Para selecionar o modelo mais adequado aos dados utilizamos o fator de Bayes[1], dado por

$$B_{10} = \frac{P(D|M_1)}{P(D|M_0)}$$

onde

$$P(D|M_k) = \int L_{M_k}(\theta|D)\pi(\theta|M_k)d\theta$$

é a esperança da função de verossimilhança dado o modelo M_k, $\pi(\theta|M_k)$ é a função de densidade *a priori* para θ sob o modelo M_k e $L_{M_k}(\theta|D)$ é a função de verossimilhança do modelo M_k, para $k = 0, 1$.

Portanto, podemos interpretar o fator de Bayes como sendo uma medida da evidência a favor de um dos modelos considerados com relação ao outro. Kass e Raftery (1995), sugerem interpretar o fator de Bayes através de uma calibragem, dividindo os possíveis valores do fator de Bayes em quatro intervalos e considerando 2 vezes o logaritmo do fator de Bayes, conforme a Tabela 3.

Tabela 3: Calibragem do fator de Bayes.

B_{10}	$2\log(B_{10})$	Evidências a favor do modelo M_1
1 a 3	0 a 2	Fraca
3 a 20	2 a 6	Moderada
20 a 150	6 a 10	Forte
> 150	> 10	Muito Forte

Fonte: Kass e Raftery (1995)

Na Tabela 3, no primeiro intervalo (1 a 3), a evidência a favor do modelo M_1 é fraca. No segundo intervalo (3 a 20), a evidência a favor do modelo M_1 aumenta, favorecendo sua escolha. No terceiro intervalo (20 a 150), a escolha do modelo M_1 pode ser feita com mais confiança, pois há uma forte evidência a seu favor. E no quarto intervalo (>150), a escolha do modelo M_1 deve ser feita.

[1] Para maiores detalhes ver Kass e Raftery 1995 ou Missão 2004.

Na análise da expressão gênica, selecionamos um modelo, M_0 ou M_1, para as medidas de níveis de expressão de um determinado gene g, $g = 1, 2, ..., G$, utilizando o fator de Bayes

$$B_{10} = \frac{P(D_c, D_t|M_1)}{P(D|M_0)}$$

onde,

$$\begin{aligned} P(D_c, D_t|M_1) &= \int_0^\infty \int_{-\infty}^\infty \int_0^\infty \int_{-\infty}^\infty L_{M_1}\left(\mu_{gc}, \sigma^2_{gc}, \mu_{gt}, \sigma^2_{gt}|D_{gc}, D_{gt}\right) \\ &\quad \pi(\mu_c, \sigma^2_c, \mu_t, \sigma^2_t|M_1) d\mu_c d\sigma^2_c d\mu_t d\sigma^2_t \end{aligned} \tag{3.7}$$

$$P(D|M_0) = \int_0^\infty \int_{-\infty}^\infty L_{M_0}\left(\mu_g, \sigma^2_g|D_g\right) \pi\left(\mu_g, \sigma^2_g|M_0\right) d\mu_g d\sigma^2_g. \tag{3.8}$$

O fator de Bayes é influênciado pela escolha das distribuições *a priori* (ver Kass e Raftery, 1995). Devido a isto, escolhemos os conjuntos de hiperparâmetros ϖ_0 e ϖ_1 de forma a reduzir a influência das distribuições *a priori* na escolha dos modelos M_0 e M_1. Baseados nas distribuições *a priori* conjunta (3.3) e (3.4), fazemos

$$\begin{aligned} \mu_{0c} &= \mu_{0t} = \mu_0 = 0 \\ \lambda_c + \lambda_t &= \lambda \\ \beta_c + \beta_t &= \beta \\ \alpha_c + \alpha_t + 3 &= \alpha \end{aligned}$$

para $\alpha > 3$.

3.3.1 Aproximação do Fator de Bayes

As integrais em (3.7) e (3.8) não possuem forma fechada conhecida, por estimativa utilizamos uma aproximação do fator de Bayes.

Kass e Raftery (1995) e Missão (2004) descrevem vários métodos para aproximar as integrais como em (3.7) e (3.8), aqui utilizamos a aproximação via método Monte Carlo para aproximação de integrais.

Utilizando uma notação simplificada para (3.7) e (3.8), podemos reescrevê-las como

$$P(D) = I_k = \int L\left(\theta_k|D\right)\pi\left(\theta_k\right)d\theta_k$$

e segundo Kass e Raftery (1995), I_k pode ser aproximada através do método Monte Carlo fazendo

$$\widehat{I}_k = \frac{1}{m}\sum_{i=1}^{m}\left[L(\theta_k^{(i)}|D)\right],$$

onde os $\theta_k^{(i)}$ são parâmetros gerados da distribuição *a priori* $\pi(\theta_k)$, m é a quantidade de $\theta_k^{(i)}$ gerados e $L(\theta_k^{(i)}|D)$ é o valor da função de verossimilhança, do modelo M_k, dado o valor gerado *a priori* $\theta_k^{(i)}$, para $k = 0, 1$ e $i = 1, 2, ...m$.

Para Geweke (1989), a precisão do método Monte Carlo pode ser melhorada pelo método de *Importance Sampling* que consiste em gerar $\theta_k^{(i)}$, para $i = 1, 2, ..., m$, de uma densidade $\pi^*\left(\theta_k\right)$ ponderada através de pesos w_i. Dessa forma I_k pode ser aproximada por

$$\widehat{I}_k = \frac{\sum_{i=1}^{m} w_i L(\theta_k^{(i)}|D)}{\sum_{i=1}^{m} w_i} \tag{3.9}$$

onde $w_i = \frac{\pi\left(\theta_k^{(i)}\right)}{\pi^*\left(\theta_k^{(i)}\right)}$.

Considerando

$$\pi^*(\theta_k) = \pi\left(\theta_k|D\right) = \frac{L\left(\theta_k|D\right)\pi\left(\theta_k\right)}{P\left(D\right)} = \frac{L\left(\theta_k|D\right)\pi\left(\theta_k\right)}{I_k}, \tag{3.10}$$

obtemos uma amostra da densidade *a posteriori* de $\theta_k|D$.

Substituindo (3.10) em (3.9), temos que

$$\hat{I}_k = \frac{\sum_{i=1}^{m}\frac{\pi\left(\theta_k^{(i)}\right)}{\pi^*\left(\theta_k^{(i)}\right)}L\left(\theta_k^{(i)}|D\right)}{\sum_{i=1}^{m}\frac{\pi\left(\theta_k^{(i)}\right)}{\pi^*\left(\theta_k^{(i)}\right)}} = \frac{\sum_{i=1}^{m}\frac{\pi\left(\theta_k^{(i)}\right)I_k}{L\left(\theta_k^{(i)}|D\right)\pi\left(\theta_k^{(i)}\right)}L\left(\theta_k^{(i)}|D\right)}{\sum_{i=1}^{m}\frac{\pi\left(\theta_k^{(i)}\right)I_k}{L\left(\theta_k^{(i)}|D\right)\pi\left(\theta_k^{(i)}\right)}}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{m} I_k}{\sum_{i=1}^{m} \frac{I_k}{L\left(\theta_k^{(i)}|D\right)}} = \frac{mI_k}{I_k \sum_{i=1}^{m} \frac{1}{L\left(\theta_k^{(i)}|D\right)}} = \frac{m}{\sum_{i=1}^{m} \frac{1}{L\left(\theta_k^{(i)}|D\right)}} \\
&= \frac{m}{\sum_{i=1}^{m} \left(L\left(\theta_k^{(i)}|D\right)\right)^{-1}} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} \left(L\left(\theta_k^{(i)}|D\right)\right)^{-1}} \\
&= \left[\frac{1}{m}\sum_{i=1}^{m} \left(L\left(\theta_k^{(i)}|D\right)\right)^{-1}\right]^{-1}
\end{aligned}
$$

onde os $\theta_k^{(i)}$ são parâmetros gerados da distribuição *a posteriori* $\pi(\theta_k|D)$, m é a quantidade de $\theta_k^{(i)}$ gerados e $L(\theta_k^{(i)}|D)$ é o valor da função de verossimilhança, do modelo M_k, dado o valor gerado *a posteriori* $\theta_k^{(i)}$, para $k = 0, 1$ e $i = 1, 2, ...m$. Como decorrência das propriedades do método Monte Carlo, este resultado converge quase certamente[2] para o valor de I_k, quando $m \to \infty$.

Assim, a aproximação do fator de Bayes para a análise da expressão gênica, seguirá os seguintes passos:

1 - Gerar computacionalmente m pares de parâmetros da distribuição *a posteriori* (3.5) do modelo M_0,

$$\left[(\mu_g, \sigma_g^2)_1, ..., \left(\mu_g, \sigma_g^2\right)_m\right];$$

2 - Gerar computacionalmente m quádruplas de parâmetros da distribuição *a posteriori* (3.6) do modelo M_1,

$$\left[(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2)_1, ..., \left(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2\right)_m\right];$$

3 - Substituir, respectivamente, os parâmetros gerados nos passos 1 e 2 acima nas respectivas verossimilhanças (3.1) e (3.2), calcular o valor de $\hat{I}_k$, $k = 0, 1$, e calcular o valor, aproximado, do fator de Bayes, dado por

$$\hat{B}_{10} = \frac{\hat{I}_1}{\hat{I}_0};$$

[2]Para detalhes sobre convergência quase certa, ver por exemplo Singer e Leite (1999).

4 - Considerar o modelo M_1 como sendo o selecionado se $\hat{B}_{10} > 1$. Caso contrário considerar o modelo M_0 como o selecionado;

Portanto, se o modelo M_1 for selecionado, temos evidências para níveis de expressão diferentes. Caso o modelo M_0 seja selecionado, não temos evidência níveis de expressão diferentes.

3.3.2 Simulação

Desenvolvemos para o fator de Bayes, um estudo de simulação similar ao realizado para o teste t, com o objetivo de verificar o seu comportamento na detecção de genes com evidências para níveis de expressão diferentes, quando consideramos diferentes afastamentos na média e na variância (ou ambos) das medidas de tratamento com relação às medidas de controle.

Os hiperparâmetros utilizados foram selecionados de forma a obtermos distribuições *a priori* pouco informativas. Para isto, consideramos os hiperparâmetros das distribuições *a priori*, dos parâmetros do modelo M_0, como sendo

$$\mu_0 = 0;\ \lambda = 0,01;\ \alpha = 3,1 \text{ e } \beta = 2,\ \beta = 1 \text{ e } \beta = 0,5.$$

Os valores diferentes para β foram utilizados para verificarmos a sensibilidade do fator de Bayes na detecção dos genes com evidências para diferença com relação a escolha dos hiperparâmetros.

Para obtermos uma distribuição *a priori* similar do ponto de vista do fator de Bayes para os parâmetros do modelo M_1, fazemos

$$\begin{aligned} \mu_{0c} &= \mu_{0t} = \mu_0 = 0 \\ \lambda_c &= \lambda_t = \frac{\lambda}{2} \\ \alpha_c &= \alpha_t = \frac{\alpha - 3}{2} \\ \beta_c &= \beta_t = \frac{\beta}{2}. \end{aligned} \tag{3.11}$$

Para gerar os dados de expressão gênica consideramos o mesmo procedimento descrito na seção 2.1 do capítulo 2. Isto é, uma amostra controle com o logaritmo dos níveis

de expressão provenientes de um distribuição normal, $N(\mu_c, \sigma_c^2)$, e uma amostra tratamento, com o logaritmo dos níveis de expressão provenientes de uma distribuição normal, $N(\mu_t, \sigma_t^2)$, onde $\mu_t = \mu_c + \delta$ e $\sigma_t^2 = (\gamma\sigma_c)^2$.

Para cada variação de δ e γ aplicamos o fator de Bayes para detectar os genes (simulados) que apresentam evidências para níveis de expressão diferentes.

Para aproximação do fator de Bayes, utilizamos $m = 10000$, ou seja, geramos 10000 pares de parâmetros

$$\left[\left(\mu_g, \sigma_g^2\right)_1, ..., \left(\mu_g, \sigma_g^2\right)_{10000}\right]$$

da distribuição *a posteriori* (3.5) do modelo M_0 e 10000 quádruplas de parâmetros

$$\left[\left(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2\right)_1, ..., \left(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2\right)_{10000}\right]$$

da distribuição *a posteriori* (3.6) do modelo M_1.

As Tabelas 4, 5 e 6 mostram as quantidades de genes (simulados) detectados com evidências para níveis de expressão diferentes, pelo fator de Bayes, para cada variação δ e γ (ou ambos) considerados, para $\beta = 2$, $\beta = 1$ e $\beta = 0,5$, respectivamente. Cada linha mostra as quantidades detectadas, quando consideramos γ fixo e variamos δ. Cada coluna mostra as quantidades detectadas, quando consideramos δ fixo e variamos γ.

Para γ fixo, cada linha das Tabelas 4, 5 e 6, e a partir de $\delta = 0$, aumentando ou diminuindo o valor de δ, o fator de Bayes detecta uma quantidade maior de genes com evidências para níveis de expressão diferentes do que a variação δ anterior. Isto nos mostra que o fator de Bayes é sensível á variação na média, detectando os genes com evidências para níveis de expressão diferentes.

Para δ fixo, cada coluna das Tabelas 5 e 6, e a partir de $\gamma = 0,25$ aumentando o valor de γ, o fator de Bayes detecta uma quantidade igual ou maior que a variação γ anterior. Isto nos mostra que o fator de Bayes é sensível tanto com relação a variação na média quanto com relação a variação na variância. Isto é, o fator de Bayes, para $\beta = 1$ e $\beta = 0,5$, detecta evidências para níveis de expressão diferentes tanto com relação a variação na média, quanto com relação a variação na variância, ou ambos. O mesmo não acontecendo na simulação realizada para o teste t.

Porém, na Tabela 4 ($\beta = 2$), para $\delta = \pm 1$, aumentando o valor de γ a quantidade

de genes detectados com evidências para diferença diminui. Isto possivelmente acontece devido a distribuição *a priori* utilizada para a variância ser não informativa, como pode ser observado pelos valores apresentados na Tabela 7, que mostra uma estatística descritiva de 10000 valores gerados (utilizando a linguagem R) da distribuição *a priori* $IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$ para a variância do modelo M_0, com $\alpha = 3,1$ e $\beta = 2$, $\beta = 1$ e $\beta = 0,5$. Dos 10000 valores gerados, com $\beta = 2$, distribuição *a priori* $IG\left(\frac{3,1}{2}, \frac{2}{2}\right)$ para a variância do modelo M_0, a média é 1,811 e a variância é 43,777. Ou seja, a utilização desta distribuição *a priori* para a variância do modelo M_0, está sendo não informativa. Com $\beta = 1$, distribuição *a priori* $IG\left(\frac{3,1}{2}, \frac{1}{2}\right)$ para a variância do modelo M_0, a média é 0,906 e a variância é 10,944 e com $\beta = 0,5$, distribuição *a priori* $IG\left(\frac{3,1}{2}, \frac{0,5}{2}\right)$ para a variância do modelo M_0, a média é 0,453 e a variância é 2,736. Ou seja, para estas duas distribuições *a priori* para a variância, temos mais informação *a priori* para a variância do modelos M_0, comparada a $IG\left(\frac{3,1}{2}, \frac{2}{2}\right)$, o que reflete nos resultados obtidos nas Tabelas 5 e 6.

Portanto, temos que o fator de Bayes é sensível aos hiperparâmetros das distribuições *a priori* para a variância dos modelos M_0 e M_1. Logo, devemos ter certa atenção e alguma informação para determinarmos os valores destes hiperparâmetros para obtermos resultados satisfatórios.

Uma forma de determinar os valores dos hiperparâmetros, seria baseados na opinião de um especialista da área genética. Se não temos a opinião do especialista podemos utilizar métodos empíricos (para detalhes sobre métodos empíricos para analise da expressão gênica ver, Efron *et al.*, 2001 e Dahl, 2002).

Dessa forma, se temos informação para definirmos os valores dos hiperparâmetros das distribuições *a priori*, acreditamos que a utilização do fator de Bayes para a análise da expressão gênica, comparada com os resultados obtidos com o estudo de simulação realizado para o teste t proposto por Baldi e Long (2001), se mostra um método estatístico complementar para se obter resultados satisfatórios.

Pois enquanto o teste t se mostra com um melhor desempenho na detecção de diferença de médias com variâncias razoavelmente estáveis ($\gamma \leq 1$), o fator de Bayes se mostra com um melhor desempenho na detecção de diferença de médias quando esta está acompanhada de aumento na variância ($\gamma > 1$).

Tabela 4: Quantidades detectadas com evidência para diferença, pelo fator de Bayes com $\beta = 2$.

	δ						
γ	-1,0	-0,8	-0.5	0,0	0,5	0,8	1,0
0,25	992	672	07	0	17	669	983
1	988	743	52	0	75	732	980
4	968	839	407	89	404	839	969
9	967	901	711	499	727	912	975
16	973	943	865	774	881	944	979

Tabela 5: Quantidades detectadas com evidência para diferença, pelo fator de Bayes com $\beta = 1$.

	δ						
γ	-1,0	-0,8	-0.5	0,0	0,5	0.8	1,0
0,25	895	266	0	0	02	278	886
1	896	414	08	0	12	394	895
4	897	663	226	36	230	681	914
9	923	825	566	326	567	821	930
16	945	898	772	658	789	901	948

Tabela 6: Quantidades detectadas com evidência para diferença, pelo fator de Bayes com $\beta = 0,5$.

	δ						
γ	-1,0	-0,8	-0.5	0,0	0,5	0.8	1,0
0,25	711	95	0	0	01	123	701
1	765	209	02	0	04	227	753
4	837	552	128	22	136	525	833
9	891	740	473	244	481	759	894
16	923	852	713	584	717	866	925

Tabela 7: Estatística descritiva dos valores gerados das distribuições *a priori*.

Valor β	D. *a priori*	Min	Média	Var	Q. 0,25	Med.	Q. 0,75	Max
$\beta = 2$	$IG(\left(\frac{3,1}{2}; 1\right)$	0,080	1,811	43,777	0,476	0,815	8,584	419,787
$\beta = 1$	$IG\left(\frac{3,1}{2}; 0,5\right)$	0,040	0,906	10,944	0,238	0,408	4,292	209,893
$\beta = 0,5$	$IG\left(\frac{3,1}{2}; 0,25\right)$	0,020	0,453	2,736	0,119	0,204	2,146	104,947

3.3.3 Aplicação

Aplicamos o fator de Bayes na análise da expressão gênica, utilizando os níveis de expressão obtidos do experimento realizado com as células da bactéria *Escherichia Coli*.

Os hiperparâmetros utilizados foram os mesmos utilizados na simulação descrita acima. Hiperparâmetros equivalentes para as distribuições *a priori* dos parâmetros do modelo M_1, foram obtidos como em (3.11). Para aproximação do fator de Bayes, utilizamos $m = 30000$.

As Figuras 4, 5 e 6 mostram as médias e variâncias controle e tratamento observadas, destacando os genes detectados com evidência para níveis de expressão diferentes, para $\beta = 2$, $\beta = 1$ e $\beta = 0,5$, respectivamente. Nestas Figuras os pontos $\bullet$ indicam os genes que não foram detectados com evidências para diferença e os sinais + indicam os genes que foram detectados com evidências para diferença.

Como observado na simulação, o fator de Bayes detecta evidências para níveis de expressão diferentes tanto com relação a variação na média quanto com relação a variação na variância (ou ambos). Isso justifica o fato do fator de Bayes ter detectado genes com evidências para diferença próximos a reta $y = x$ no gráfico das médias.

Nas aplicações do fator de Bayes, os casos em que um número maior de genes é identificado, estes contêm os genes selecionados em casos em que se identificou um número menor de genes.

Para $\beta = 2$; 43 genes foram detectados com evidências para diferença, com 5 genes (44, 273, 323, 366 e 370) com média de controle e de tratamento próximos a reta $y = x$ sendo detectados com evidências para diferença, devido a variação presente na variância. Porém, os genes 44 e 273 possuem média e variância de controle e de tratamento próximos (ver Tabela 8). Assim, acreditamos que estes genes tenham sido identificados com evidências para diferença devido a distribuição *a priori* utilizada ser não informativa, como citado

na simulação (ver Figura 4).

Para $\beta = 1$; 31 genes foram detectados com evidências para diferença, com 2 genes (323 e 370) com média de controle e de tratamento próximos à reta $y = x$ sendo detectados com evidências para diferença, devido a variação presente na variância de controle e de tratamento deste gene (ver Figura 5).

Para $\beta = 0,5$; 24 genes foram detectados com evidências para diferença, com 2 genes (323 e 370) com média de controle e de tratamento próximos a reta $y = x$ sendo detectados com evidências para diferença devido a variação presente na variância de controle e de tratamento deste gene (ver Figura 6).

Como as variâncias das medidas dos níveis de expressão dos genes das células da bactéria *Escherichia Coli* sao todas altamente concentradas no intervalo (0;0,3), podemos verificar pelos valores da Tabela 7 que dos 10000 valores gerados, quando utilizamos $\beta = 2$, o menor valor é 0,080 e o maior é 419,787 e apenas 879 (8,79%) são menores que 0,3. Ou seja, a utilização desta distribuição *a priori*, $IG\left(\frac{3,1}{2}, \frac{2}{2}\right)$ para a variância do modelo M_0, como observado na simulação, está sendo não informativa, pois poucos valores gerados pertencem ao domínio dos valores observados para variância. Conseqüentemente, está sendo não informativa para a variância do modelo M_1,devido a equivalência em (3.11). Para $\beta = 1$ e $\beta = 0,5$ temos distribuições *a priori* para a variância do modelo M_0 mais informativa do que quando utilizamos $\beta = 2$ (ver Tabela 7), o que reflete nos resultados obtidos.

Portanto, como observado na simulação, o fator de Bayes é sensível a escolha dos hiperparâmetros das distribuições *a priori*, levando à escolha de diferentes genes g, $g = 1, 2, ..., G$, com evidências para níveis de expressão diferentes. Para distribuições *a priori* com informação para as variâncias dos modelos M_0 e M_1, por exemplo, quando utilizamos $\beta = 1$ ou $\beta = 0,5$, os resultados obtidos são satisfatórios, pois os genes com média controle e tratamento distantes da reta $y = x$ e com diferença presente na variância das medidas de tratamento com relação as medidas de controle, são detectados com evidências para diferença.

Comparando os resultados obtidos com a aplicação do fator de Bayes e do teste t (ver Figuras 2 a 6 e Tabelas 2 e 8), podemos observar uma melhor performance do fator de Bayes. Pois, tanto na simulação quanto na aplicação, ao contrário do teste t, o fator

de Bayes, é capaz de detectar genes com evidências para diferença tanto com relação a variação na média quanto com relação a variação na variância, ou ambos. Isto mostra, porque o teste t não detectou o gene 323, que apresenta diferença na variância, com evidências para diferença, e o fator de Bayes detectou. Um outro exemplo é o gene 277 citado anteriormente (ver Figura 6), que apresenta diferença tanto na média quanto na variância, porém o teste t não o detectou, e o fator de Bayes o detectou com evidências para diferença.

Embora o teste t, com um nível de significância de $0,10$, tenha identificado 51 genes com evidência para diferença, portanto mais do que o fator de Bayes, somente os genes 134, 179, 233, 248, 376 e 383 foram detectados pelos dois métodos (ver Tabelas 2 e 8).

Assim, se temos informação para definirmos os valores dos hiperparâmetros (através da opinião de um especialista da área genética ou utilizando métodos empíricos), acreditamos que a utilização do fator de Bayes, com relação a utilização do teste t, forneça um acréscimo, não de quantidade, mas sim de qualidade na seleção dos genes com evidências para níveis de expressão diferentes, quando estes possuem evidências para diferença tanto com relação à média quanto com relação a variâncias das observações de tratamento com relação as observações de controle.

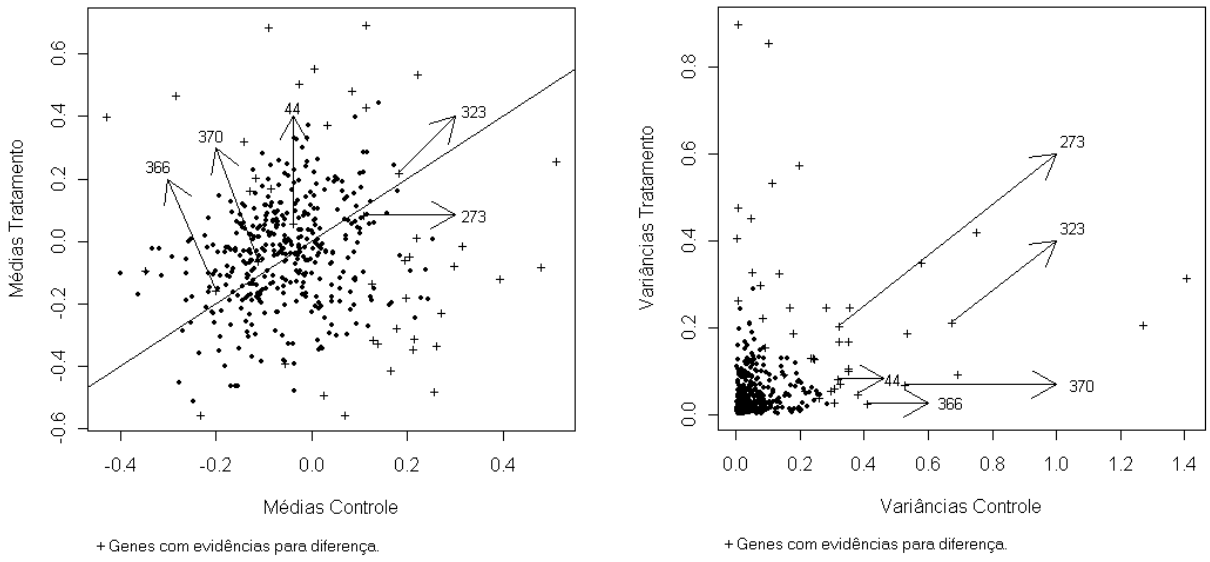

Figura 4: Médias e variâncias controle e tratamento, com $\beta = 2$.

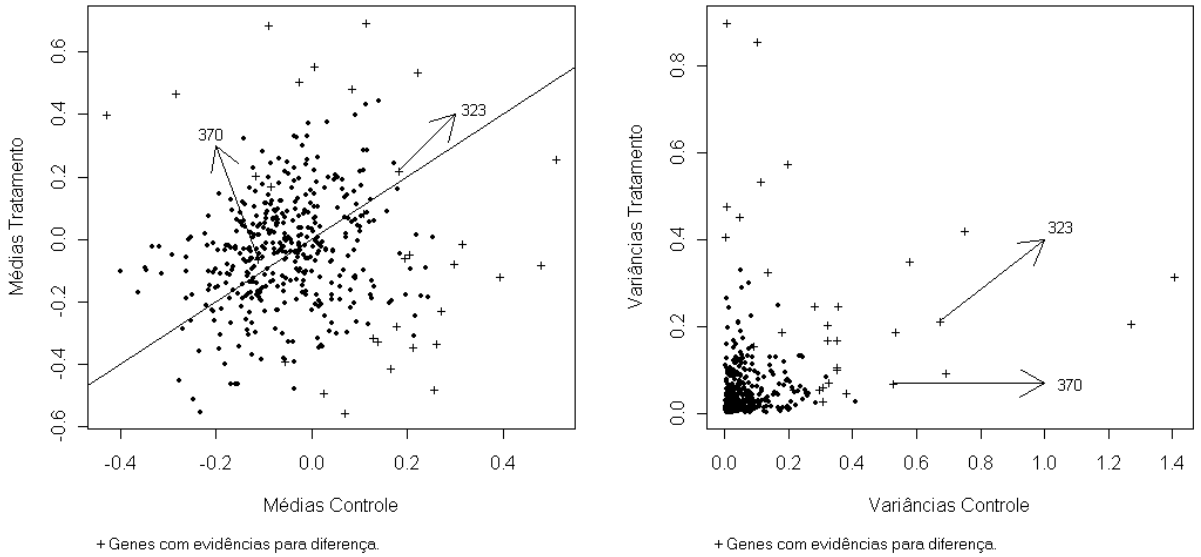

Figura 5: Médias e variâncias controle e tratamento, com $\beta = 1$.

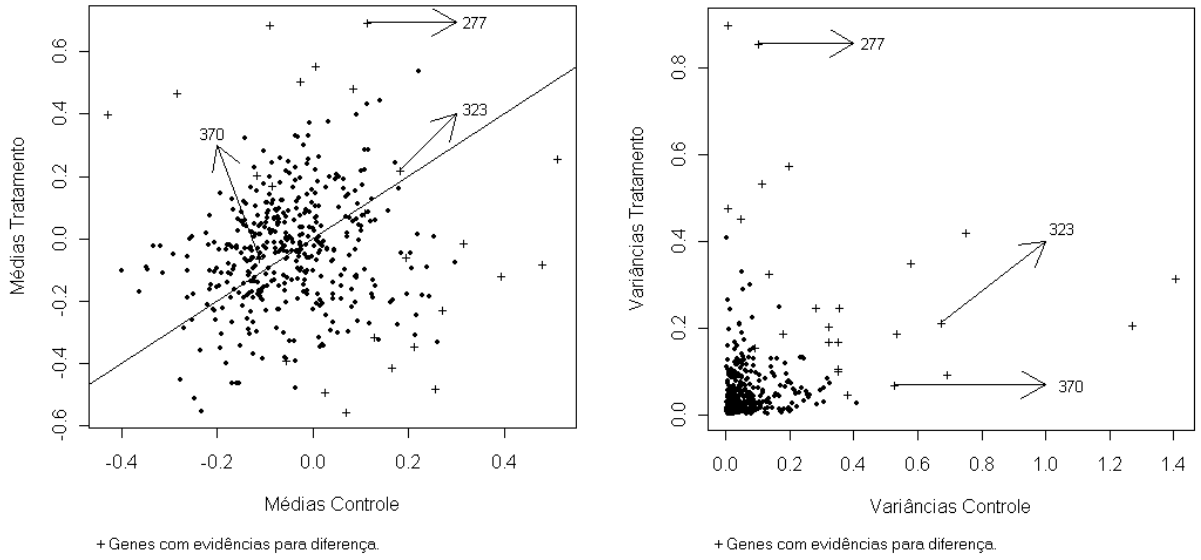

Figura 6: Médias e variâncias controle e tratamento, com $\beta = 0,5$.

A Tabela 8 mostra o número e as médias e variâncias controle e tratamento observadas, dos 43 genes detectados com evidências para níveis de expressão diferentes, pelo fator de Bayes quando utilizamos $\beta = 2$. Os genes indicados com * são os detectados quando utilizamos $\beta = 1$ e os indicados com $\circ$ são os detectados quando $\beta = 0,5$. Observe que todos os genes detectados quando utilizamos $\beta = 1$ e $\beta = 0,5$ possuem evidências para diferença, tanto com relação a média quanto com relação a variância (ou ambos).

Tabela 8: Genes detectados com evidência para diferença pelo fator de Bayes.

N^o gene	Média cont	Variância cont.	Média trat	Variância trat.
19	-0,3484	0,2444	-0,0912	0,1286
44	-0,0392	0,3195	0,0609	0,0824
*83	0,1393	0,3084	-0,3253	0,0628
92	0,2048	0,3063	-0,0447	0,0302
113	0,1252	0,2346	-0,1336	0,1320
*115	0,5121	0,3225	0,2593	0,1698
$\overset{*}{\circ}$121	0,0064	0,1121	0,5560	0,5337
$\overset{*}{\circ}$133	0,3939	0,3811	-0,1173	0,0470
*134	0,2603	0,2152	-0,3331	0,0567
$\overset{*}{\circ}$156	0,1278	0,2814	-0,3109	0,2470
*159	-0,1287	0,1667	0,1645	0,2485
$\overset{*}{\circ}$179	0,0705	0,1365	-0,5547	0,3277
*190	-0,2330	0,2416	-0,5548	0,1329
*193	0,1781	0,2976	-0,2767	0,0573
194	0,1688	0,1872	-0,1955	0,1178
$\overset{*}{\circ}$201	0,1650	0,3507	-0,4115	0,1697
$\overset{*}{\circ}$213	-0,0549	1,2703	-0,3886	0,2062
$\overset{*}{\circ}$226	-0,0850	0,3563	0,1736	0,2488
*233	0,0249	0,1770	-0,4876	0,1894
237	0,2185	0,0773	0,0134	0,2983
$\overset{*}{\circ}$247	0,4788	0,1960	-0,0782	0,5741
248	0,2133	0,1489	-0,3095	0,0939
$\overset{*}{\circ}$250	-0,1159	0,5342	0,2073	0,1889
$\overset{*}{\circ}$251	-0,4292	1,4074	0,4018	0,3153
$\overset{*}{\circ}$252	0,1951	0,6933	-0,0583	0,0950
$\overset{*}{\circ}$256	-0,0271	0,0454	0,5082	0,4525
*270	0,2217	0,0015	0,5363	0,4073

Nº gene	Média cont	Variância cont.	Média trat.	Variância trat.
*273	0,1084	0,3224	0,0846	0,2061
*°277	0,1144	0,1010	0,6958	0,8571
320	0,1129	0,0503	0,4315	0,3295
*°323	0,1827	0,6725	0,2209	0,2133
*°324	-0,0886	0,0066	0,6860	0,9007
*°328	0,0838	0,0061	0,4823	0,4766
*329	0,2984	0,3268	-0,0773	0,0730
332	-0,1419	0,0071	0,3220	0,2635
*348	0,3152	0,3509	-0,0109	0,1074
366	-0,2005	0,4087	-0,1553	0,0268
*°370	-0,1118	0,5260	-0,0625	0,0690
376	0,2174	0,0490	-0,2458	0,1898
379	0,1979	0,2576	-0,1776	0,0409
*°383	0,2566	0,0903	-0,4787	0,1561
*°385	0,2114	0,7492	-0,3441	0,4202
*°415	0,2706	0,3514	-0,2245	0,1030
*°417	-0,2840	0,5785	0,4681	0,3519

3.4 Seleção de Modelos: DIC

Considere $y_1, y_2, ..., y_n$ uma amostra observada que pode ser associada a um de dois modelos de probabilidade, que chamaremos de modelo M_0 e M_1. A cada um dos modelos temos, respectivamente, associada uma função densidade ou de probabilidade e uma função de verossimilhança,

$$f_{M_1}(y|\theta_1) \text{ e } L_{M_1}(\theta_1|y) = \prod_{i=1}^{n} f(y|\theta_1)$$

$$f_{M_2}(y|\theta_2) \text{ e } L_{M_2}(\theta_2|y) = \prod_{i=1}^{n} f(y|\theta_2).$$

onde θ_1 e θ_2 representam o(s) parâmetro(s) do modelo M_0 e M_1, respectivamente.

Nesta seção utilizamos o DIC (*Deviance Information critérion*), introduzido por Spiegelhalter *et al.*, (2002), para selecionar o modelo que melhor explica os dados observados.

Para selecionar o modelo que melhor explica os dados observados é calculado o DIC para cada um dos modelos, M_0 e M_1, e o modelo que apresentar o menor valor DIC é o modelo que melhor explica os dados observados. Assim, se temos k modelos candidatos, $M_0, M_1, ..., M_k$, devemos calcular para cada um dos k modelos o DIC e o modelo que apresentar menor valor e o modelo que melhor explica os dados observados.

O cálculo do DIC, para cada um dos modelo candidatos, é baseado no cálculo da *deviance* (denotada por $D(\theta)$), definida por

$$D(\theta) = -2 \log L(\theta|y) + 2 \log f(y) \tag{3.12}$$

onde $f(y)$ é uma função apenas dos dados.

Para comparação de modelos, utilizamos $f(y) = 1$ para todos os modelos (ver, Dempster, 1974). Assim, (3.12) é dada por

$$D(\theta) = -2 \log L(\theta|y). \tag{3.13}$$

Baseado no cálculo da *deviance* em (3.12), Spiegelhalter *et al* (2002), desenvolve o critério de seleção de modelos DIC, definido por

$$DIC = \bar{D} + P_D \tag{3.14}$$

onde $\bar{D}$ é a esperança *a posteriori* da *deviance*,

$$\bar{D} = E_{\theta|y} \left[D(\theta) \right] \tag{3.15}$$

e o termo P_D é chamado de número de parâmetros efetivos no modelo, que é a medida de complexidade do modelo, e é definido como sendo a diferença entre a esperança *a posteriori* da *deviance* e a *deviance* calculada no valor esperado dos parâmetros estimados *a posteriori*, isto é,

$$P_D = E_{\theta|y} \left[D\left(\theta\right) \right] - D\left[E_{\theta|y}\left(\theta\right) \right] = \bar{D} - D\left(\bar{\theta}\right) \tag{3.16}$$

onde $\bar{\theta}$ é a média dos valores gerados *a posteriori* para o(s) parâmetros θ.

Utilizando o método de Monte Carlo, o DIC para cada um dos modelos candidatos pode ser calculado a cada valor gerado do(s) parâmetro(s) θ, onde a cada iteração calcula-se o valor de $D(\theta)$, como em (3.13). Ao final das iterações, calcula-se a média dos valores de $D(\theta)$, obtendo-se $\bar{D}$, e subtrai-se desse valor a estimativa da *deviance* calculada, utilizando a média amostral $\bar{\theta}$ dos valores gerados *a posteriori* para o(s) parâmetro(s) θ, obtendo-se assim o valor de P_D, dado em (3.16). Logo, temos o valor DIC dado como em (3.14).

Para análise da expressão gênica, selecionamos o modelo M_0 ou M_1, definidos na seção 3.1, para um determinado gene g, $g = 1, 2, ..., G$, utilizando o critério DIC.

Para cada um dos modelos, M_0 ou M_1 e um determinado gene g, temos as verossimilhanças dadas em (3.1) e (3.2). Logo as *deviances*, como em (3.13), são dadas por

$$\begin{aligned} D_{M_0}(\mu_g, \sigma_g^2) &= -2\log L_{M_0}(\mu_g, \sigma_g^2|D) \\ &\propto -2\log\left[\left(\sigma_g^2\right)^{-\frac{n}{2}}\exp\left\{-\frac{n(\bar{x}_g-\mu_g)^2+(n-1)s_g^2}{2\sigma_g^2}\right\}\right] \\ &\propto n\log\left(\sigma_g^2\right)+\frac{n(\bar{x}_g-\mu_g)^2+(n-1)s_g^2}{\sigma_g^2} \end{aligned} \tag{3.17}$$

$$\begin{aligned} D_{M_1}(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2) &= -2\log L_{M_1}(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2|D_c, D_t) \\ &= -2\log\left[L_c(\mu_{gc}, \sigma_{gc}^2, |D_c)L_t(\mu_{gt}, \sigma_{gt}^2|D_t)\right] \\ &\propto n_c\log\left(\sigma_{gc}^2\right)+\frac{n_c(\bar{x}_{gc}-\mu_{gc})^2+(n_c-1)s_{gc}^2}{\sigma_{gc}^2} \\ &\quad +n_t\log\left(\sigma_{gt}^2\right)+\frac{n_t(\bar{x}_{gt}-\mu_{gt})^2+(n_t-1)s_{gt}^2}{\sigma_{gt}^2}. \end{aligned} \tag{3.18}$$

O valor DIC associado aos modelos M_0 e M_1, como em (3.14), são dados por

$$\mathrm{DIC}_{M_0} = \bar{D}_{M_0} + P_{D_{M_0}} \tag{3.19}$$

$$\mathrm{DIC}_{M_1} = \bar{D}_{M_1} + P_{D_{M_1}} \tag{3.20}$$

onde

$$P_{D_{M_0}} = \bar{D}_{M_0} - D\left(\bar{\mu}_g, \bar{\sigma}_g^2\right) \text{ e } P_{D_{M1}} = \bar{D}_{M_1} - D\left(\bar{\mu}_{gc}, \bar{\sigma}_{gc}^2, \bar{\mu}_{gt}, \bar{\sigma}_{gt}^2\right)$$

com $\bar{\mu}_g$ e $\bar{\sigma}^2_g$ sendo a média dos valores gerados para μ_g e σ^2_g da distribuição *a posteriori* (3.5) e $\bar{\mu}_{gc}, \bar{\sigma}^2_{gc}, \bar{\mu}_{gt}, \bar{\sigma}^2_{gt}$ sendo a média dos valores gerados para $\mu_{gc}, \sigma^2_{gc}, \mu_{gt}, \sigma^2_{gt}$ da distribuição *a posteriori* (3.6).

Assim, o cálculo do DIC para análise da expressão gênica, sob os modelos M_0 e M_1, segue os seguintes passos:

1 - Gerar computacionalmente da distribuição *a posteriori* (3.5) do modelo M_0,

$$\left[\left(\mu_g, \sigma^2_g\right)_1, ..., \left(\mu_g, \sigma^2_g\right)_m\right]$$

m pares de parâmetros;

2 - Gerar computacionalmente da distribuição *a posteriori* (3.6) modelo M_1,

$$\left[\left(\mu_g, \sigma^2_g\right)_1, ..., \left(\mu_g, \sigma^2_g\right)_m\right]$$

m quádruplas de parâmetros;

3 - Substituir, respectivamente, os parâmetros gerados nos passos 1 e 2 acima nas respectivas *deviances* (3.17) e (3.18);

4 - calcular o valor de $\bar{D}_{M_0}$, $\bar{D}_{M_1}$, $P_{D_{M_0}}$ e $P_{D_{M_1}}$;

5 - calcular o valor de DIC_{M_0} e DIC_{M_1}, como em (3.19) e (3.20).

Portanto, se o modelo M_1 apresentar menor valor DIC, $\text{DIC}_{M_1} < \text{DIC}_{M_0}$, temos evidências para níveis de expressão diferentes. Caso o modelo M_0 tenha menor valor DIC, $\text{DIC}_{M_0} < \text{DIC}_{M_1}$, não temos evidências para níveis de expressão diferentes.

3.4.1 Simulação

Desenvolvemos para o critério DIC, o mesmo estudo de simulação realizado para o fator de Bayes e para o teste t, com o objetivo de verificar seu comportamento na detecção de genes com evidências para níveis de expressão diferentes, quando consideramos diferentes afastamentos na média e/ou na variância das medidas de tratamento com relação às medidas de controle.

Para gerar os dados de expressão gênica, consideramos o mesmo procedimento descrito para o fator de Bayes (seção 3.3) e para o teste t (seção 2.1).

Os hiperparâmetros utilizados para as distribuições *a priori*, dos parâmetros do modelo M_0, foram os mesmos utilizados na simulação e realizada para o fator de Bayes. Para determinarmos os hiperparâmetros das distribuições *a priori,* dos parâmetros do modelo M_1, utilizamos as mesmas restrições dadas em (3.11).

Os diferentes valores de β foram utilizados para verificarmos a sensibilidade do critério DIC na detecção de evidências para diferença com relação a escolha dos hiperparâmetros.

Para cada variação considerada em δ e γ (ou ambos) e cada valor de β aplicamos o critério DIC para selecionar o modelo M_0 ou M_1 e detectar os genes (simulados) que apresentam evidências para níveis de expressão diferentes.

Para o cálculo do DIC para cada um dos modelos, M_0 e M_1, utilizamos $m = 10000$, ou seja, geramos 10000 pares de parâmetros

$$\left[\left(\mu_g, \sigma_g^2\right)_1, ..., \left(\mu_g, \sigma_g^2\right)_{10000}\right]$$

da distribuição *a posteriori* (3.5) do modelo M_0 e 10000 quádruplas de parâmetros

$$\left[\left(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2\right)_1, ..., \left(\mu_{gc}, \sigma_{gc}^2, \mu_{gt}, \sigma_{gt}^2\right)_{10000}\right]$$

da distribuição *a posteriori* (3.6) do modelo M_1 e calculamos o valor DIC_{M_0} e DIC_{M_1} como em (3.19) e (3.20).

As Tabelas 9, 10 e 11 mostram as quantidades de genes (simulados) detectados com evidências para níveis de expressão diferentes, pelo critério DIC, para cada variação de δ e γ (ou ambos) considerados, para $\beta = 2$, $\beta = 1$ e $\beta = 0,5$, respectivamente. Cada linha mostra as quantidades detectadas, quando consideramos γ e fixo e δ variando. Cada coluna mostra as quantidades detectadas, quando consideramos δ fixo e γ variando.

Para γ fixo, cada linha das Tabelas 9, 10 e 11, a partir de $\delta = 0$ aumentado ou diminuindo o valor de δ, o critério DIC detecta uma quantidade maior de genes com evidências para diferença do que a variação δ anterior. Ou seja, o critério DIC como o fator de Bayes é sensível a variação na média detectando os genes com evidências para níveis de expressão diferentes.

Para δ fixo, cada coluna das Tabelas 10 e 11, aumentando o valor de γ o critério DIC detecta uma quantidade maior de genes com evidências para diferença. Ou seja, o

critério DIC também é sensível a variação na variância. Isto nos mostra que o critério DIC, como o fator de Bayes, para $\beta = 1$ e $\beta = 0,5$, detecta evidências para níveis de expressão diferentes tanto com relação a variação na média quanto com relação a variação na variância (ou ambos), das medidas de tratamento com relação as medidas de controle.

Porém, na Tabela 9 ($\beta = 2$), para $\delta = \pm 1$, aumentando o valor de γ a quantidade de genes detectados com evi- dências para diferença diminui. Isto possivelmente acontece devido a distribuição *a priori* utilizada para a variância dos modelo M_0 e M_1 ser não informativa, como citado na simulação realizada para o fator de Bayes.

Portanto, temos que o critério DIC, como o fator de Bayes, é sensível a escolha dos hiperparâmetros das distribuições *a priori*, levando a detecção de diferentes genes com evidências para diferença. Dessa forma, devemos ter certa atenção e alguma informação para determinarmos os valores dos hiperparâmetros para obtermos resultados satisfatórios. Uma forma de determinarmos os valores dos hiperparâmetros, seria baseados na opinião de um especialista da área genética ou utilizar métodos empíricos.

Os resultados obtidos com o critério DIC e com o fator de Bayes, são semelhantes (ver Tabelas 4, 5, 6, 9, 10 e 11). Dessa forma, se temos informação para definirmos os valores dos hiperparâmetros das distribuições *a priori*, acreditamos que a utilização do critério DIC para a análise da expressão gênica, como o fator de Bayes, se mostra um método estatístico complementar para se obter resultados satisfatórios, comparado com os resultados obtidos com o teste t proposto por Baldi e Long (2001).

Tabela 9: Quantidades detectadas com evidência para diferença, pelo critério DIC com $\beta = 2$.

	δ						
γ	-1,0	-0,8	-0.5	0,0	0,5	0.8	1,0
0,25	996	727	18	0	19	728	994
1	988	780	66	0	79	778	987
4	975	852	419	65	414	851	978
9	973	898	699	455	713	914	975
16	971	938	858	756	869	940	973

Tabela 10: Quantidades detectadas com evidência para diferença, pelo critério DIC com $\beta = 1$.

	δ						
γ	-1,0	-0,8	-0.5	0,0	0,5	0.8	1,0
0,25	872	224	0	0	01	240	881
1	874	379	05	0	08	373	885
4	895	650	202	22	207	664	909
9	916	803	535	267	541	813	926
16	941	879	741	605	758	892	943

Tabela 11: Quantidades detectadas com evidência para diferença, pelo critério DIC com $\beta = 0,5$.

	δ						
γ	-1,0	-0,8	-0.5	0,0	0,5	0.8	1,0
0,25	586	46	0	0	0	65	580
1	689	148	02	0	01	180	687
4	807	503	110	09	112	496	812
9	872	705	421	179	418	727	873
16	909	835	662	511	675	838	915

3.4.2 Aplicação

Aplicamos o critério DIC na análise da expressão gênica, utilizando as medidas de níveis de expressão obtidos do experimento realizado com as células da bactéria *Escherichia Coli*.

Os hiperparâmetros utilizados foram os mesmo utilizados na simulação.

Nas aplicações do critério DIC, como nas aplicações do fator de Bayes, os casos em que uma quantidade maior de genes é detectada, este contém os genes identificados em casos que se detectou uma quantidade menor de genes.

As Figura 7, 8 e 9 mostram as médias e variâncias observadas para as medidas de controle e de tratamento, destacando os genes detectados com evidências para diferença.

Para $\beta = 2$; 38 genes foram detectados com evidências para diferença, com 3 genes

(273, 323, 370) com média de controle e de tratamento próximos a reta $y = x$ sendo detectados com evidência para diferença, devido a variação presente na variância das medidas de tratamento com relação as medidas de controle. Porém, como na aplicação do fator de Bayes, o gene 273 que possui média e variância de controle e de tratamento próximos foi detectado com evidências para diferença. Acreditamos que este gene (273) tenha sido identificado com evidência para diferença, devido a distribuição *a priori* utilizada para a variância da situação de controle e tratamento, deste gene, ser não informativa, como citado naS simulações realizadas para o fator de Bayes e para o critério DIC.

Para $\beta = 1$; 25 genes foram detectados com evidência para diferença, com 2 genes (323, 370) com média de controle e de tratamento próximo a reta $y = x$ sendo detectado com evidências para diferença devido a variação presente na variância das medidas de tratamento com relação as medidas de controle, destes genes (ver Figura 8).

Para $\beta = 0,5$; 18 genes foram detectados com evidência para diferença, com 1 gene (323) com média de controle e de tratamento próximo a reta $y = x$ sendo detectado com evidências para diferença devido a variação presente na variância das medidas de tratamento com relação as medidas de controle, deste gene (ver Figura 9).

Como observado na simulação, o critério DIC como o fator de Bayes, é sensível a escolha dos hiperparâmetros, levando a detecção de diferentes genes com evidências para níveis de expressão diferentes. Para distribuições *a priori* com informação para a variância das medidas de níveis de expressão da situação de controle e de tratamento, por exemplo, quando utilizamos $\beta = 1$ e $\beta = 0,5$, os resultados obtidos são satisfatórios, pois os genes com média de controle e de tratamento distantes da reta $y = x$ e com diferença na variância das medidas de tratamento com relação as medidas de controle, foram detectados com evidências para diferença. Exemplos são os genes 277 e 323 já citados, que apresentam respectivamente diferença tanto na média quanto na variância e somente na variância.

Todos os genes detectados com evidências para níveis de expressão diferentes pelo critério DIC, também foram detectados pelo fator de Bayes.

Se comparado ao fator de Bayes, o critério DIC detectou uma quantidade menor de genes com evidências para diferença. Porém essa diminuissão foi de uma forma positiva, pois somente os genes com média de controle e de tratamento próximos a reta $y = x$ deixaram de ser detectados (ver Figuras 4 a 9 e Tabelas 8 e 12).

Com relação aos detectados pelo teste t, os genes 134, 179, 233, 248 e 383 foram detectados pelo critério DIC (ver Tabelas 2 e 12).

Assim, se temos informação para determinarmos os valores dos hiperparâmetros das distribuições *a priori*, acreditamos que a utilização do critério DIC e do fator de Bayes, forneça um acréscimo de qualidade na detecção dos genes com evidências para níveis de expressão diferentes, com relação aos detectados pelo teste t, quando temos evidências para diferença de médias, entre tratamento e controle, acompanhada de aumento na variância das observações de tratamento com relação a variância das observações de controle.

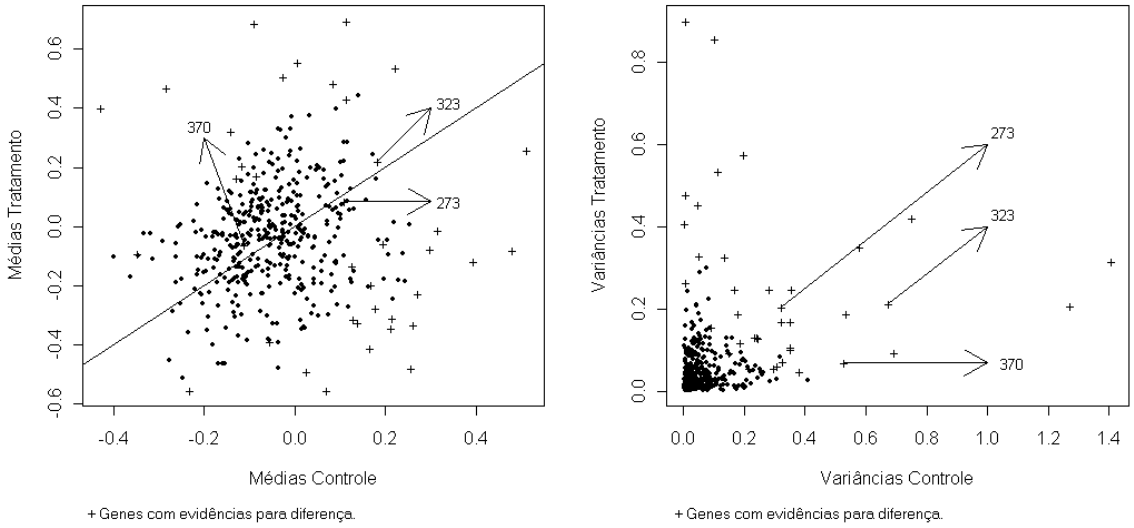

Figura 7: Médias e Variâncias conmtrole e tratamento, com $\beta = 2$.

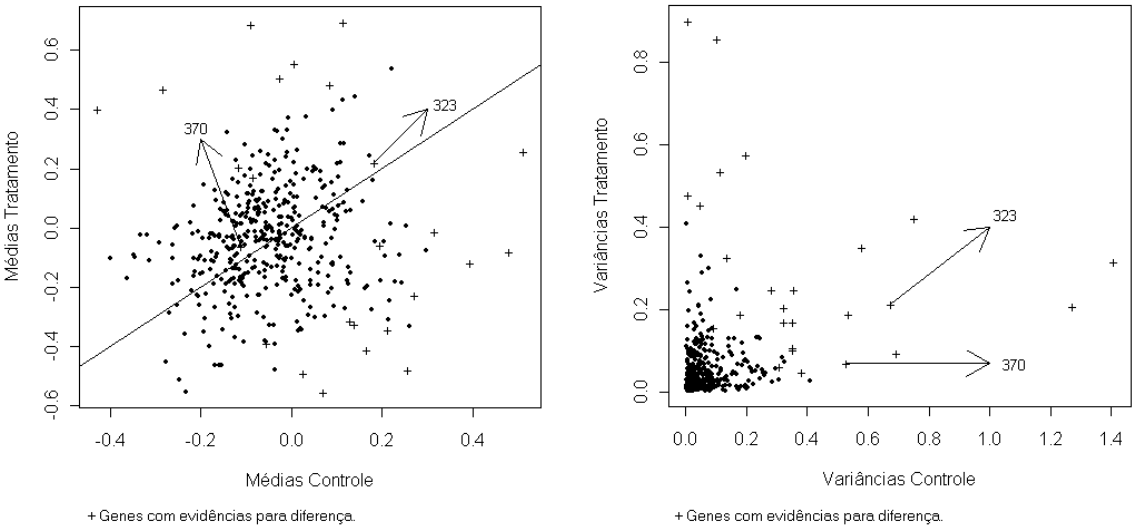

Figura 8: Médias e Variâncias conmtrole e tratamento, com $\beta = 1$.

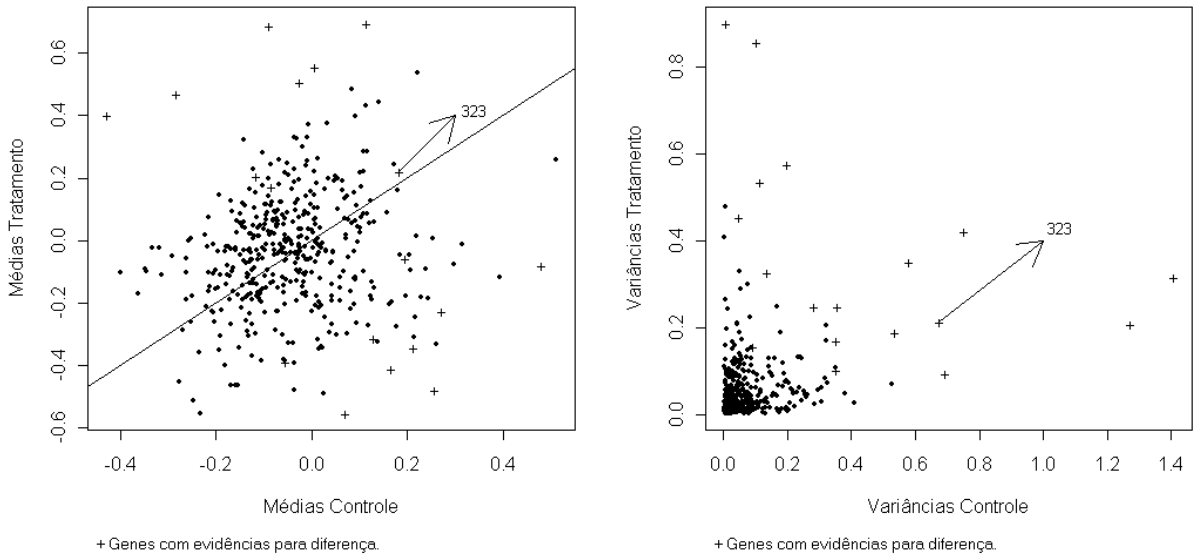

Figura 9: Médias e Variâncias conmtrole e tratamento, com $\beta = 0,5$.

A Tabela 12 mostra o número e as médias e variâncias controle e tratamento observadas, dos 38 genes detectados com evidências para níveis de expressão diferentes, pelo DIC. Os genes indicados com * são os detectados quando utilizamos $\beta = 1$ e os indicados com $\circ$ são os detectados quando $\beta = 0,5$. Observe que todos os genes detectados quando utilizamos $\beta = 1$ e $\beta = 0,5$ possuem evidências para diferença, tanto com relação a média quanto com relação a variância (ou ambos).

Tabela 12: Genes detectado com evidência para diferença.

N^o gene	Média cont	Variância cont.	Média trat	Variância trat.
19	-0,3484	0,2444	-0,0912	0,1286
44	-0,0392	0,3195	0,0609	0,0824
*83	0,1393	0,3084	-0,3253	0,0628
113	0,1252	0,2346	-0,1336	0,1320
*115	0,5121	0,3225	0,2593	0,1698
$^*_\circ$121	0,0064	0,1121	0,5560	0,5337
$^*_\circ$133	0,3939	0,3811	-0,1173	0,0470
*134	0,2603	0,2152	-0,3331	0,0567
$^*_\circ$156	0,1278	0,2814	-0,3109	0,2470
159	-0,1287	0,1667	0,1645	0,2485
$^*_\circ$179	0,0705	0,1365	-0,5547	0,3277

N^o gene	Média cont	Variância cont.	Média trat	Variância trat.
190	-0,2330	0,2416	-0,5548	0,1329
193	0,1781	0,2976	-0,2767	0,0573
194	0,1688	0,1872	-0,1955	0,1178
**201	0,1650	0,3507	-0,4115	0,1697
**213	-0,0549	1,2703	-0,3886	0,2062
**226	-0,0850	0,3563	0,1736	0,2488
*233	0,0249	0,1770	-0,4876	0,1894
**247	0,4788	0,1960	-0,0782	0,5741
248	0,2133	0,1489	-0,3095	0,0939
**250	-0,1159	0,5342	0,2073	0,1889
**251	-0,4292	1,4074	0,4018	0,3153
**252	0,1951	0,6933	-0,0583	0,0950
**256	-0,0271	0,0454	0,5082	0,4525
270	0,2217	0,0015	0,5363	0,4073
273	0,1084	0,3224	0,0846	0,2061
**277	0,1144	0,1010	0,6958	0,8571
320	0,1129	0,0503	0,4315	0,3295
**323	0,1827	0,6725	0,2209	0,2133
**324	-0,0886	0,0066	0,6860	0,9007
*328	0,0838	0,0061	0,4823	0,4766
329	0,2984	0,3268	-0,0773	0,0730
*348	0,3152	0,3509	-0,0109	0,1074
*370	-0,1118	0,5260	-0,0625	0,0690
**383	0,2566	0,0903	-0,4787	0,1561
**385	0,2114	0,7492	-0,3441	0,4202
**415	0,2706	0,3514	-0,2245	0,1030
**417	-0,2840	0,5785	0,4681	0,3519

Capítulo 4

Abordagem bayesiana não paramétrica: Processo Dirichlet

Modelos incorporando o processo Dirichlet *a priori* têm iniciado um importante e recente desenvolvimento de aplicações bayesianas. A flexibilidade deste modelo, aliado ao desenvolvimento dos métodos computacionais, tem possibilitado sua utilização em diversas áreas. Neste capítulo, discutimos e ilustramos o uso do processo Dirichlet e do modelo de misturas de processos Dirichlet na inferência bayesiana não paramétrica e/ou semi-paramétrica. Posteriormente, aplicamos o modelo de misturas de processos Dirichlet na análise da expressão gênica.

4.1 Introdução

Considere $Y_1, Y_2, ..., Y_n$ variáveis aleatórias independentes e identicamente distribuídas ($i.i.d$), definidas em um espaço Ω com distribuição de probabilidades G. Se G for definida em um contexto paramétrico, então G será conhecida e caracterizada por um conjunto finito de parâmetros $\theta \in \Theta$, onde Θ representa o espaço paramétrico. Dentro de uma abordagem bayesiana, definimos um modelo de probabilidades *a priori* para θ e fazemos inferências baseados em sua distribuição *a posteriori*.

Se eliminamos a suposição de que G é caracterizada por um conjunto finito de parâmetros $\theta \in \Theta$, G será desconhecida, e na abordagem bayesiana, entramos na abordagem conhecida como inferência bayesiana não paramétrica. Na inferência bayesiana o termo

não paramétrico sugere um modelo abrangente com infinitos parâmetros (ver, Walker *et al.*, 1999). Neste caso, ao invez de especificarmos um modelo de probabilidades *a priori* para θ (como no modelo paramétrico) especificamos um modelo de probabilidades *a priori* para G.

Na abordagem bayesiana não paramétrica, é desejável que a distribuição *a priori* tenha algumas propriedades de modo a facilitar e viabilizar o desenvolvimento das análises (ver Antoniak, 1974). São elas:

1) A classe de distribuições *a priori* deve ser analiticamente tratável em três aspectos:

i) ser possível determinar a distribuição *a posteriori*, dada uma amostra,

ii) ser possível determinar os valores esperados das funções de perda simples,

iii) a classe de distribuições *a priori* deve ser fechada, ou seja, a distribuição *a posteriori* deve pertencer à mesma classe da distribuição *a priori*;

2) A classe de distribuições *a priori* deve ser capaz de expressar qualquer informação ou conhecimento;

3) A classe de distribuições *a priori* deve ser parametrizada de forma a produzir uma interpretação clara da informação inserida.

Satisfazer todas essas propriedades simultaneamente é uma tarefa difícil, pois em geral, a realização de uma delas implica no comprometimento das outras.

Diversos autores, entre eles Kraft e van Eeden (1964), Kraft (1964) e Dubins e Freedmam (1966), descrevem processos que satisfazem a segunda propriedade mas não satisfazem a primeira e a terceira propriedade.

Ferguson (1973) define um processo chamado processo Dirichlet, que satisfaz a primeira e a terceira propriedade e é ligeiramente deficiente com relação a segunda propriedade.

Para Ibrahim (2002) o processo Dirichlet é o principal processo em inferência bayesina não paramétrica.

Em 1974, Antoniak fez uma extensão do processo Dirichlet criando o que chamamos de modelo de misturas de processos Dirichlet.

A utilização do modelo de misturas de processos Dirichlet é indicada em análise de modelos com mistura de distribuições. Neste caso a modelagem considera que o número de componentes k da mistura é um número aleatório.

O uso do modelo de mistura de processos Dirichlet se torna computacionalmente viável

com o desenvolvimento de métodos de amostragem baseados em cadeias de *Markov*. Métodos baseados em *Gibbs Sampling* podem ser facilmente implementados para modelos com distribuições *a priori* conjugadas, porém quando distribuições *a priori* não conjugadas são utilizadas, o que é apropriado em muitos contextos, a aplicação do método *Gibbs Sampling* requer a realização de uma integração numérica como em (4.20), muitas vezes, difícil de ser realizada. Porém, já há diversos algoritmos descritos capazes de solucionar este problema, e consequentemente obter amostras das distribuições condicionais. Entre eles podemos citar: West, Müller e Escobar (1994) com uma aproximação *Monte Carlo* para a integral, MacEachern e Müller (1998) desenvolveram um método para obter amostras das distribuições condicionais utilizando um conjunto de variáveis auxiliares para o conjunto de parâmetros (algoritmo "*no gaps*") e Neal (1998) com a utilização do *Metropolis-Hastings*.

4.2 Distribuição de Dirichlet

A distribuição de Dirichlet pode ser obtida através da distribuição $Gama$, $Gama(\alpha, \beta)$, onde

- α é o parâmetro de forma, $\alpha \geq 0$,

- β é o parâmetro de escala, $\beta > 0$,

- para $\alpha = 0$, temos que $Gama(\alpha, \beta) \equiv 0$

- para $\alpha > 0$ e uma variável aleatória $Z > 0$, tal que $Z \sim Gama(\alpha, \beta)$, temos que sua função densidade de probabilidade é dada por

$$f(z|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}$$

com

$$E(Z) = \frac{\alpha}{\beta} \text{ e } Var(Z) = \frac{\alpha}{\beta^2}.$$

Considere $Z_1, Z_2, ..., Z_k$ variáveis aleatórias independentes com $Z_j \sim Gama(\alpha_j, 1)$, e $\alpha_j > 0$ para algum $j = 1, 2, ..., k$. A distribuição de Dirichlet, $k-1$ dimensional, com

parâmetros $\alpha_1, \alpha_2, ..., \alpha_k$ é a distribuição de $\mathbf{Y} = (Y_1, Y_2, ..., Y_k)$, tal que,

$$Y_j = \frac{Z_j}{\sum_{j=1}^{k} Z_j}$$

para $j = 1, 2, ..., k$. Sua função densidade de probabilidade é dada por

$$f\left(y_1, y_2, ..., y_{k-1} | \alpha_1, \alpha_2, ..., \alpha_k\right) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma\left(\alpha_j\right)} \prod_{j=1}^{k} y_j^{\alpha_j - 1} I_{\Delta}(y_1, y_2, ..., y_{k-1}), \tag{4.1}$$

onde $\Delta = \left\{(y_1, y_2, ..., y_{k-1}) : 0 < y_j < 1,\ j = 1, 2, ..., k-1 \text{ e } \sum_{j=1}^{k-1} y_j < 1\right\}$ e $y_k = 1 - \sum_{j=1}^{k-1} y_j$ com $\alpha_0 = \sum_{j=1}^{k} \alpha_j$.

É importante notar que:

- os valores assumidos pelas variáveis aleatórias Y_j, $j = 1, 2, ..., k$, estão no intevalo $[0, 1]$ e que a soma de seus valores é igual a 1, justificando o $k - 1$ dimensional da distribuição de Dirichlet.

- a distribuição de Dirichlet é uma generalização da distribuição beta.

Se $\mathbf{X} = (X_1, X_2, ..., X_n)$ é um vetor aleatório definido em um espaço discreto $\chi = \{\chi_1, \chi_2, ..., \chi_k\}$ e $\theta = (\theta_1, \theta_2, ..., \theta_k)$, tal que $P\left(X_i = x_j\right) = \theta_j$, onde x_j representa o valor que a variável aleatória X_i assume quando é associada ao evento χ_j, então

$$(X_1, X_2, ..., X_n) \sim Multinomial\left(n, \theta_1, \theta_2, ..., \theta_k\right)$$

e a distribuição *a priori* natural para $\theta = (\theta_1, \theta_2, ..., \theta_k)$ é a distribuição de Dirichlet com parâmetros $\alpha_1, \alpha_2, ..., \alpha_k$,

$$\theta = (\theta_1, \theta_2, ..., \theta_k) \sim Dir\left(\alpha_1, \alpha_2, ..., \alpha_k\right). \tag{4.2}$$

para $i = 1, 2, ..., n$ e $j = 1, 2, ..., k$.

Em particular, a distribuição de cada θ_j, do vetor $\theta = (\theta_1, \theta_2, ..., \theta_k)$, segue uma distribuição beta

$$\theta_j \sim beta\,(\alpha_j, \alpha_0 - \alpha_j)$$

para $j = 1, 2, ..., k$.

Assim, temos que a esperança e a variância de cada θ_j é a esperança e a variância da distribuição beta com parâmetros $(\alpha_j, \alpha_0 - \alpha_j)$,

$$E\,(\theta_j) = \frac{\alpha_j}{\alpha_0} \text{ e } Var\,(\theta_j) = \frac{\alpha_j\,(\alpha_0 - \alpha_j)}{\alpha_0^2\,(\alpha_0 + 1)} \tag{4.3}$$

para $j = 1, 2, ..., k$.

Se

$$(\theta_1, \theta_2, ..., \theta_k) \sim Dir\,(\alpha_1, \alpha_2, ..., \alpha_k)\,,$$

a moda para um parâmetro θ_j é dada por

$$Moda\,(\theta_j) = \frac{\alpha_j - 1}{\alpha_0 - k}$$

e a covariância entre dois parâmetros, θ_i e θ_j, para $i \neq j$, é dada por

$$Cov\,(\theta_i, \theta_j) = -\frac{\alpha_i\ \alpha_j}{\alpha_0^2\,(\alpha_0 + 1)}.$$

Utilizando o teorema de Bayes, temos que a distribuição *a posteriori* de θ em (4.2), é dada por

$$(\theta_1, \theta_2, ..., \theta_k | X_1 = x_1, X_2 = x_2, ..., X_k = x_k) \sim Dir\,(\alpha_1 + x_1, \alpha_2 + x_2, ..., \alpha_k + x_k) \tag{4.4}$$

onde $x_j \geq 0$, $\sum_{j=1}^{k} x_j = n$ e $\sum_{j=1}^{k} \theta_j = 1$.

Como podemos notar, a distribuição de Dirichlet é conjugada com relação a distribuição multinomial.

Se

$$(\theta_1, \theta_2, ..., \theta_k) \sim Dir\,(\alpha_1, \alpha_2, ..., \alpha_k)$$

e agrupamos alguns θ_j, obtemos uma nova distribuição de Dirichlet com parâmetros

iguais a soma dos parâmetros correspondentes aos θ_j´s agrupados. Ou seja, dado $r = \{r_1, r_2, ..., r_t\}$, com r_j inteiro e $j = 1, 2, ..., t$, tal que $0 < r_1 < r_2 < ... < r_t = k$, então

$$\left(\sum_{j=1}^{r_1} \theta_j, \sum_{j=r_1+1}^{r_2} \theta_j, ..., \sum_{j=r_{t-1}+1}^{r_t} \theta_j \right) \sim Dir \left(\sum_{j=1}^{r_1} \alpha_j, \sum_{j=r_1+1}^{r_2} \alpha_j, ..., \sum_{j=r_{t-1}+1}^{r_t} \alpha_j \right). \quad (4.5)$$

4.3 Processo Dirichlet (PD)

A extensão da distribuição de Dirichlet para espaços contínuos é conhecida como processo Dirichlet.

O processo Dirichlet foi introduzido por Ferguson (1973) como um método para resolução de problemas de natureza não paramétrica na abordagem bayesiana.

Baseado em Ferguson (1973), o processo Dirichlet tem a seguinte definição:

- Seja Ω um espaço amostral e $\mathbf{A}$ uma $\sigma-$álgebra de subconjuntos de Ω,
- α uma medida positiva ($\alpha > 0$) sobre o espaço $(\Omega, \mathbf{A})$,
- $B_1, B_2, ..., B_k$ uma partição mensurável de Ω,
- $(G(B_1), G(B_2), ..., G(B_k))$ um vetor de probabilidades sobre a partição $B_1, B_2, ..., B_k$.

Dizemos que uma medida de probabilidade aleatória G sobre $(\Omega, \mathbf{A})$ é um processo Dirichlet sobre o espaço $(\Omega, \mathbf{A})$, com parâmetro α, denotado por $G \sim PD(\alpha)$, se para qualquer partição $B_1, B_2, ..., B_k$ de Ω o vetor de probabilidades $(G(B_1), G(B_2), ..., G(B_k))$ segue uma distribuição de Dirichlet com parâmetros $(\alpha(B_1), \alpha(B_2), ..., \alpha(B_k))$,

$$(G(B_1), G(B_2), ..., G(B_k)) \sim Dir\,(\alpha(B_1), \alpha(B_2), ..., \alpha(B_k)). \quad (4.6)$$

Se G é uma função de distribuição desconhecida, isto é, $G : \mathbb{R} \to [0, 1]$, tal que

- G é não decrescente,
- G é continua a direita e tem limite pela esquerda,
- $\lim_{x \to -\infty} G(x) = 0$ e $\lim_{x \to \infty} G(x) = 1$.

Então dado G existe uma medida de probabilidade, g, induzida por G em $(\mathbb{R}, \mathbf{ß})$, onde $\mathbf{ß}$= σ-álgebra de subconjuntos de Borel em $\mathbb{R}$, tal que

- $g\,((a, b]) = G(b) - G(a)$, com $a \in \mathbb{R}$, $b \in \mathbb{R}$ e $-\infty \leq a < b < \infty$;

- $g\left((-\infty, b]\right) = G(b)$;

- $g\left(\{a\}\right) = G(a) - \lim_{x \to a^-} G(x)$.

O problema de atribuir a G uma distribuição de probabilidades *a priori* passa a ser um problema de "aleatorização"da g. Para isto, supomos que G possui distribuição *a priori* definida por um processo Dirichlet com parametro α em $(\mathbb{R}, \mathbf{ß})$, $G \sim PD(\alpha)$, significando que $g \sim PD(\alpha)$. Isto é, para todo $k \geq 1$ e toda partição $B_1, B_2, ..., B_k$ disjunta de $(\mathbb{R}, \mathbf{ß})$, temos que

$$(g(B_1), g(B_2), ..., g(B_k)) \sim Dir\left(\alpha(B_1), \alpha(B_2), ..., \alpha(B_k)\right),$$

ou seja, a metodologia consite em dizer que os efeitos aleatórios, presentes na variável de interesse, se distribuem segundo uma função de distribuição desconhecida G, de tal sorte que sua medida de probabilidade induzida g seja distribuidas segundo um processo Dirichlet.

A definição do PD dada acima fica evidente se consideramos, por exemplo, $\Omega = \mathbb{R}$, $B_1, B_2, ..., B_k$ como sendo intervalos disjuntos, tal que $B_j \in \mathbf{A}$, $\mathbf{A} = \mathbf{ß}^1$ e $g = f_{N(\mu,\sigma^2)}(B_j)$, com μ e σ^2 desconhecidos. Como cada $g(B_j)$ é a probabilidade do intervalo B_j na distribuição normal, para $j = 1, 2, ..., k$, e estas probabilidades são desconhecidas e discretas, uma distribuição de probabilidades *a priori* adequada para o vetor $(g(B_1), g(B_2), ..., g(B_k))$ é a distribuição de Dirichlet.

Assim temos que:

- cada $g\left(B_j\right)$ é uma variável aleatória definida no intervalo $(0, 1)$, para todo $B_j \in \mathbf{A}$ e $j = 1, 2, ..., k$;

- as propriedades da distribuição de Dirichlet, descritas nas seção 4.2, se estendem para o processo Dirichlet.

Portanto, se $g \sim PD\left(\alpha\right)$ e $A \in \mathbf{A}$, então como em (4.3), temos que

$$E\left[g\left(A\right)\right] = \frac{\alpha(A)}{\alpha\left(\Omega\right)} \tag{4.7}$$

e

$$Var\left[g\left(A\right)\right] = \frac{\alpha\left(A\right)\left[\alpha\left(\Omega\right) - \alpha\left(A\right)\right]}{\alpha^2\left(\Omega\right)\left[\alpha\left(\Omega\right) + 1\right]}. \tag{4.8}$$

[1] **σ-álgebra de subconjuntos de Borel: Classe de intervalos (a,b], $-\infty \leq a < b < \infty$. Tais intervalos são chamados de conjuntos "básicos"ou "cilíndricos"de ß.**

Como em (4.5), se $r = \{r_1, r_2, ..., r_t\}$ com r_j, $j = 1, 2, ..., t$, inteiros tal que $0 < r_1 < r_2 < ... < r_t = k$ e agrupamos alguns B_j,

$$\left(\bigcup_{j=1}^{r_1} B_j, \bigcup_{j=r_1+1}^{r_2} B_j, ..., \bigcup_{j=r_{t-1}+1}^{r_t} B_j \right)$$

temos que

$$\left(g\left(\bigcup_{j=1}^{r_1} B_j \right), ..., g\left(\bigcup_{j=r_{t-1}+1}^{r_t} B_j \right) \right) = \left(\sum_{j=1}^{r_1} g(B_j), ..., \sum_{j=r_{t-1}+1}^{r_t} g(B_j) \right)$$

e, obtemos uma nova distribuição de Dirichlet com parâmetros iguais a soma dos parâmetros correspondentes aos $B_{j's}$ agrupados, ou seja

$$\left(\sum_{j=1}^{r_1} g(B_j), ..., \sum_{j=r_{t-1}+1}^{r_t} g(B_j) \right) \sim Dir \left(\sum_{j=1}^{r_1} \alpha\left(B_j\right), ..., \sum_{j=r_{t-1}+1}^{r_t} \alpha\left(B_j\right) \right).$$

Considerando $\alpha = Mg_0$, onde g_0 é uma função densidade de probabilidade de uma função de distribuição paramétrica G_0 especificada e M, $M > 0$, é um parâmetro de precisão.

Portanto, se $g \sim PD\left(\alpha = Mg_0\right)$, então podemos reescrever (4.6) como

$$\left(g(B_1), g(B_2), ..., g(B_k)\right) \sim Dir\left(Mg_0(B_1), Mg_0(B_2), ..., Mg_0(B_k)\right). \tag{4.9}$$

Como em (4.7) e (4.8), dado $A \in \mathbf{A}$, temos que

$$E\left[g\left(A\right)\right] = \frac{Mg_0\left(A\right)}{Mg_0\left(\Omega\right)} = \int_A dG_0\left(A\right) = g_0(A) \tag{4.10}$$

e

$$\begin{aligned} Var\left[g\left(A\right)\right] &= \frac{Mg_0\left(A\right)\left[M - Mg_0\left(A\right)\right]}{\left[Mg_0\left(\Omega\right)\right]^2\left(Mg_0\left(\Omega\right) + 1\right)} = \frac{Mg_0\left(A\right)\left[M - Mg_0\left(A\right)\right]}{M^2\left(M+1\right)} \\ &= \frac{g_0\left(A\right)\left[1 - g_0\left(A\right)\right]}{M+1}, \end{aligned} \tag{4.11}$$

ou seja, g_0 e M definem a esperança e a variância do PD. G_0 é conhecida como função de

distribuição base e Mg_0 como medida base do PD.

Temos portanto, que G representa uma função de distribuição desconhecida que segue um processo Dirichlet e que se comporta, em média, segundo uma função de distribuição paramétrica G_0 conhecida; e M é uma medida positiva e finita que indica a distância entre a função de distribuição aleatória G e sua média G_0. Devido a isto, M é comumente interpretado como controlador da variabilidade das medidas de probabilidades aleatórias G sobre G_0 (ver Walker *et al.*, 1999).

Uma motivação para se trabalhar com o PD *a priori* é a obtenção direta de sua distribuição *a posteriori*. Suponha $y_1, y_2, ..., y_n$ uma amostra aleatória proveniente de uma função de distribuição G, desconhecida, e que se comporta *a priori* segundo um processo Dirichlet, $g \sim PD(Mg_0)$. Isto é, dado $B_1, B_2, ..., B_k$ uma partição mensurável de Ω, com $B_j's \in \mathbf{A}$ e $\alpha = Mg_0$, então

$$(g(B_1), g(B_2), ..., g(B_k)) \sim Dir\left(Mg_0(B_1), Mg_0(B_2), ..., Mg_0(B_k)\right).$$

A distribuição *a posteriori* é dada por,

$$G|y_1, ..., y_n \sim PD\left(Mg_0(B_1) + \sum_{i=1}^{n} I_{B_1}(y_1), ..., Mg_0(B_k) + \sum_{i=1}^{n} I_{B_k}(y_i)\right) \tag{4.12}$$

onde, $I_{B_j}(y_i)$ é uma função indicadora que coloca massa 1 no ponto y_i, para $i = 1, 2, ..., n$ e $j = 1, 2, ..., k$. Isto é, para cada $B_1, B_2, ..., B_k$, definidos arbitrariamente, tal que $B_j \in \mathbf{A}$,

$$I_{B_j}(y_i) = \begin{cases} 1, \text{ se } y_i \in B_j \\ 0, \text{ se } y_i \notin B_j \end{cases}$$

e

$$\sum_{i=1}^{n} I_{B_j}(y_i) = \#\{y_i, 1 \leq i \leq n : y_i \in B_j\}$$

onde # representa a quantidade de $y_i's$ que pertencem a partição B_j, para $i = 1, 2, .., n$ e $j = 1, 2, ..., k$.

Note que (4.12) pode ser deduzida utilizando o mesmo procedimento mostrado em (4.4), ou seja, a atualização *a posteriori* é feita pela quantidade de observações pertencentes aos respectivos conjuntos $B_j's$. Muitos autores, tais como Ibrahim (1998), sugerem

"construir" as partições de modo que cada um dos B_j´s considerados contenha pelo menos uma das observações y_i, $i = 1, 2, ..., n$.

Por simplicidade, muitos autores escrevem (4.12) apenas como

$$G|y_1, y_2, ..., y_n \sim PD\left(Mg_0 + \sum_{i=1}^{n} I_{B_j}(y_i)\right). \tag{4.13}$$

para $j = 1, 2, ..., k$.

Considere um conjunto $A = (-\infty, y]$, logo

$$\sum_{i=1}^{n} I_A(y_i) = \sum_{i=1}^{n} I_{((-\infty,y])}(y_i) = \#\{y_i : y_i \leq y\} = nG_n(y)$$

onde, G_n é a função de distribuição empírica, definida por

$$G_n(y) = \frac{\#\ de\ y_i's\ \leq y\ na\ amostra}{n}$$

para $i = 1, 2, ..., n$.

A distribuição *a posteriori* de G, dados $y_1, y_2, ..., y_n$ é um PD com parâmetros

$$Mg_0 + \sum_{i=1}^{n} I_{y_i} = Mg_0 + nG_n.$$

Assim, podemos escrever (4.13), como

$$G|y_1, y_2, ..., y_n \sim PD(Mg_0 + nG_n).$$

Note que

$$Mg_0(\Omega) + nG_n(\Omega) = (Mg_0 + nG_n)(\Omega) = M + n$$

e que a esperança *a posteriori*, como em (4.10), é dada por

$$\begin{aligned} E[g(A)|y_1, ..., y_n] &= \frac{g_0(A) + nG_n(A)}{Mg_0(\Omega) + nG_n(\Omega)} = \frac{Mg_0(A) + nG_n(A)}{M + n} \\ &= \frac{M}{M+n} g_0(A) + \frac{n}{M+n} G_n(A). \end{aligned} \tag{4.14}$$

Dessa forma, temos de (4.14) que

- se $M \to \infty$, então

$$\frac{M}{M+n} = \frac{1}{1+\frac{n}{M}} \to 1$$

e

$$\frac{n}{M+n} = 1 - \frac{M}{M+n} \to 0,$$

ou seja, um grande "peso" é dado a função de distribuição base G_0 e um pequeno "peso" é dado as observações;

- se $M \to n$, então

$$\frac{M}{M+n} \to \frac{1}{2}$$

e

$$\frac{n}{M+n} = 1 - \frac{M}{M+n} \to \frac{1}{2},$$

ou seja, o mesmo "peso" é dado a função de distribuição base G_0 e as observações;

- se $M \to 0$, então

$$\frac{M}{M+n} \to 0$$

e

$$\frac{n}{M+n} \to 1,$$

ou seja, um pequeno "peso" é dado a função de distribuição base G_0 e um grande "peso" é dado as observações.

Exemplo: Sejam $y_1, y_2, ..., y_6$ uma amostra aleatória proveniente de uma funçao de distribuição G, desconhecida, definida no conjunto $\Omega = \mathbb{R}^+$ com $\mathbf{A} = \mathbf{ß}$. Suponha que G tenha distribuição *a priori* definida por um PD, $G \sim PD(Mg_0)$.

Considere o conjunto de dados

$$y_1 = 0,8;\ y_2 = 1;\ y_3 = 1,4;\ y_4 = 2,1;\ y_5 = 1,6 \text{ e } y_6 = 3,2$$

e os seguintes parâmetros para o PD

$$M = 1 \text{ e } G_0 = F_{\exp(1)} = 1 - e^{-y} \text{ e } g_0 = e^{-y}.$$

Note que estamos dando um maior peso para as observações, $\frac{n}{M+n} = \frac{6}{1+6} = \frac{6}{7} \cong 0,86$ contra $\frac{M}{M+n} = \frac{1}{1+6} = \frac{1}{7} \cong 0,14$ dado a função de distribuição base G_0.

Para construirmos as partições $B_{j's}$, nos baseamos nas observações, de modo que cada uma das partiçoes contenha pelo menos uma das observações y_i, $i = 1, 2, ..., 6$.

$$\begin{aligned} B_1 &= \{y : 0 < y \leq 1\}, \\ B_2 &= \{y : 1 < y \leq 2\}, \\ B_3 &= \{y : 2 < y \leq 3\}, \\ B_4 &= \{y : y > 3\}. \end{aligned} \tag{4.15}$$

Logo,

$$\begin{aligned} g(B_1) &= G(1) - G(0), \\ g(B_2) &= G(2) - G(1), \\ g(B_3) &= G(3) - G(2) \\ g(B_4) &= 1 - G(3) = 1 - (g(B_1) + g(B_2) + g(B_3)) \end{aligned}$$

e

$$g(B_1) + g(B_2) + g(B_3) + g(B_4) = 1.$$

Como $G \sim PD(Mg_0)$ e $M = 1$, temos que

$$(g(B_1), g(B_2), g(B_3), g(B_4)) \sim PD(g_0(B_1), g_0(B_2), g_0(B_3), g_0(B_4))$$

onde

$$\begin{aligned} g_0(B_1) &= G_0(1) - G_0(0) = G_0(1) = 1 - e^{-1} = 0,6321 \\ g_0(B_2) &= G_0(2) - G_0(1) = 1 - e^{-2} - 1 + e^{-1} = e^{-1} - e^{-2} = 0,2325 \\ g_0(B_3) &= G_0(3) - G_0(2) = 1 - e^{-3} - 1 + e^{-2} = e^{-2} - e^{-3} = 0,0855 \\ g_0(B_4) &= 1 - (g_0(B_1) + g_0(B_2) + g_0(B_3)) = 1 - (0,63212 + 0,23254 + 0,08555) \\ &= 0,0497. \end{aligned}$$

Portanto, *a priori*

$$(g(B_1), g(B_2), g(B_3), g(B_4)) \sim PD(0,6321; 0,2325; 0,0855; 0,0497).$$

Note que diferentes partições do espaço amostral produzem diferentes definições de $g(B_j)$ e consequentemente diferentes valores para os hiperparâmetros $g_0(B_j)$.

Como em (4.12), a distribuição *a posteriori* de $G(B_1), ..., G(B_6)$ dados $y_1, y_2, ..., y_6$ é dada por

$$(g(B_1), ..., g(B_6))|y_1, ..., y_6 \sim PD\left(Mg_0(B_1) + \sum_{i=1}^{n} I_{y_i}(B_1), ..., Mg_0(B_6) + \sum_{i=1}^{n} I_{y_i}(B_6)\right)$$

onde,

$$\begin{aligned}
\sum_{i=1}^{n} I_{y_i}(B_1) &= nG_n(1) - nG_n(0) \\
&= 6\frac{\#\ de\ y_j \leq 1\ na\ amostra}{6} - 6\frac{\#\ de\ y_j\ n \leq 0\ a\ amostra}{6} = 2 \\
\sum_{i=1}^{n} I_{y_i}(B_2) &= nG_n(2) - nG_n(1) \\
&= 6\frac{\#\ de\ y_j \leq 2\ na\ amostra}{6} - 6\frac{\#\ de\ y_j \leq 1\ na\ amostra}{6} = 2 \\
\sum_{i=1}^{n} I_{y_i}(B_3) &= nG_n(3) - nG_n(2) \\
&= 6\frac{\#\ de\ y_j \leq 3\ na\ amostra}{6} - 6\frac{\#\ de\ y_j \leq 2\ na\ amostra}{6} = 1 \\
\sum_{i=1}^{n} I_{y_i}(B_4) &= nG_n(\infty) - nG_n(3) \\
&= 6 - 6\frac{\#\ de\ y_j \leq 3\ na\ amostra}{6} = 6 - 5 = 1
\end{aligned}$$

Logo, a distribuição *a posteriori* de $g(B_1), ..., g(B_6)$ dados $y_1, y_2, ..., y_6$ é dada por

$$\begin{aligned}
(g(B_1), ..., g(B_6))|y_1, ..., y_6 &\sim PD(0,6321 + 2; 0,2325 + 2; 0,0855 + 1; 0,0497 + 1) \\
&\sim PD(2,6321; 2,2325; 1,0855; 1,0497).
\end{aligned}$$

Um aspecto interessante do PD é que ele pode ser utilizado, *a priori*, com sucesso na análise de dados com misturas de distribuições.

Em um modelo com mistura de distribuições paramétricas com um número finito de componentes, é assumido que cada observação y_i, $i = 1, 2, ..., n$, é proveniente de uma das k (k fixo) distribuições. Por exemplo, em um modelo com mistura de distribuições normais, com variâncias conhecidas, é assumido que cada observação y_i, $i = 1, 2, ..., n$, pode vir de uma das k distribuições possíveis, parametrizadas por k diferentes médias. E um aspecto neste tipo de análise é a determinação do número k de componentes da mistura.

Com a utilização do PD *a priori*, temos uma parametrização que torna a distribuição marginal de y_i, $i = 1, 2, ..., n$, por exemplo, uma mistura de distribuições normais com o número de componentes k sendo aleatório (ver Neal 1998 e MacEachern e Müller 1998).

Esta modelagem é conhecida como modelo de misturas de processos Dirichlet.

4.4 Modelo de Misturas de Processos Dirichlet

O modelo de misturas de processos Dirichlet (MPD) foi introduzido por Antoniak (1974). Recentemente, com o desenvolvimento de métodos computacionais, o modelo MPD vem sendo utilizado como método prático por diversos pesquisadores, tais como Escobar (1994), Escobar e West (1995), MacEachern e Müller (1998), Neal (1998) entre outros.

O modelo de mistura de processos Dirichlet (MPD), é essencialmente um modelo bayesiano hierárquico que considera que os dados $y_1, y_2, ..., y_n$, são provenientes de uma mistura de distribuições, isto é, $y_i \sim f(\theta_i)$ e que remove a suposição de uma distribuição paramétrica para θ_i, substituindo-a por uma distribuição qualquer G que possui distribuição *a priori* definida por um processo Dirichlet com parâmetro de concentração M e distribuição base G_0 (isto é, medida base Mg_0). Isto dá origem ao seguinte modelo

hierárquico com três estágios,

$$\begin{aligned} &estágio\ 1 \quad : \quad y_i|\theta_i \overset{ind.}{\sim} f(\theta_i) \\ &estágio\ 2 \quad : \quad \theta_i|G \overset{i.i.d}{\sim} G \\ &estágio\ 3 \quad : \quad G|M, G_0 \sim PD(Mg_0), \end{aligned} \tag{4.16}$$

para $i = 1, 2, ..., n$.

Note que a especificação em (4.16) dá origem a um modelo bayesiano semi-paramétrico, onde a distribuição paramétrica é dada no estágio 1 e a distribuição não paramétrica nos estágios 2 e 3.

Como visto na seção 4.3, podemos interpretar o parâmetro de precisão M como sendo o controlador da variabilidade das medidas de G com relação a G_0. Logo, se $M \to \infty$ um grande "peso" é dado a distribuição base G_0 e assim G_0 (que é uma distribuição paramétrica) será a distribuição *a priori* para todos os $\theta_{i's}$, para $i = 1, 2, ..., n$, e retornamos à inferência bayesiana usual.

Abaixo comparamos a modelagem MPD com a paramétrica usual, utilizando um modelo com mistura de distribuições normais com médias μ_i e variâncias constantes e igual a σ^2. Observe que a segunda e a terceira linha do modelo paramétrico exerce a mesma função que as linhas 2, 3 e 4 do modelo MPD, que é especificar a distribuição *a priori* para μ_i e σ^2. Os hiperparâmetros μ_0, σ_0^2, a, b e M podem ser fixos ou aleatórios, dependendo da aplicação e do interesse do pesquisador.

Modelo Paramétrico	Modelo MPD
$y_i\|\mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$	$y_i\|\mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$
$\mu_i \sim N(\mu_0, \sigma_0^2)$	$\mu_i, \sigma^2\|G \sim G$
$\sigma^2 \sim IG(a, b)$	$G\|M, G_0 \sim PD(Mg_0)$
	$G_0 \sim N(\mu_0, \sigma_0^2)\, IG(a, b)$

para $i = 1, 2, ..., n$.

BlacKwell e MacQueen (1973) mostram que se $\theta_1, \theta_2, ..., \theta_n \sim G$ e $G \sim PD(Mg_0)$, então $\theta_1, \theta_2, ..., \theta_n$, no limite, em geral segue o modelo da urna de Polya, podendo ser

representado da seguinte maneira

$$(\theta_i \,|\theta_1, \theta_2, ..., \theta_{i-1}, M, G_0) \sim \frac{M}{M+i-1} G_0 + \frac{1}{M+i-1} \sum_{j=1}^{i-1} I_{\theta_j} \tag{4.17}$$

com função densidade de probabilidade, dada por

$$(\theta_i|\theta_1, \theta_2, ..., \theta_{i-1}, M, G_0) = \frac{M}{M+i-1} g_0 + \frac{1}{M+i-1} \sum_{\substack{j=1 \\ j \neq i}}^{i-1} I_{\theta_j}$$

para $i = 1, 2, ..., n$.

Assim, podemos dizer que a distribuição *a priori* para $\theta_1, \theta_2, ..., \theta_n$ é uma mistura de distribuições, com uma parte contínua, gerada segundo uma distribuição paramétrica G_0 e uma parte discreta, que repete os valores das componentes anteriores. Com isso, temos que alguns valores de $\theta_1, \theta_2, ..., \theta_n$ podem ser iguais e cada observação $y_1, y_2, ..., y_n$ associada a um mesmo θ_i é modelada por uma densidade comum $f(\theta_i)$, $i = 1, 2, ..., n$.

Considere (4.17), porém agora, não condicionado sobre $\theta_1, \theta_2, ..., \theta_{i-1}$, mas sobre todos os θ*s* indexados de 1 a n, exceto o $i - ésimo$ (denotamos este vetor por $\theta_{-i} = (\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n)$). Isto pode ser feito porque as amostras obtidas de um PD são permutáveis, significando que a distribuição conjunta das variáveis não dependem da ordem que elas são consideradas (ver Neal 1998). Assim, considerando que θ_i é a última observação, temos que a distribuição condicional de θ_i é

$$(\theta_i \,|\theta_{-i}, M, G_0) \sim \frac{M}{M+n-1} G_0 + \frac{1}{M+n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} I_{\theta_j}$$

cuja função densidade de probabilidade, dada por

$$(\theta_i \,|\theta_{-i}, M, G_0) = \frac{M}{M+n-1} g_0 + \frac{1}{M+n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} I_{\theta_j} \tag{4.18}$$

para $i = 1, 2, ..., n$.

Combinando (4.18) com a função de verossimilhança $L(\theta_j|y_i) = f\,(y_i|\theta_j)$ (baseada apenas na observação y_i, isto é, a função de verossimilhança é o valor da densidade f com

parâmetro(s) θ_j no ponto y_i), temos que a densidade condicional para θ_i, é dada por

$$(\theta_i|\theta_{-i}, y_i, M, G_0) = bMf\,(y_i;\theta_i)\,g_0(\theta_i) + b\sum_{\substack{j=1\\ j\neq i}}^{n} f\,(y_i|\theta_j)\,I_{\theta_j} \tag{4.19}$$

onde b é uma constante normalizadora apropriada, para $i = 1, 2, ..., n$.

Considere

$$q_{0i} = \int g_0(\theta_i) f\,(y_i;\theta_i)\,d\theta_i. \tag{4.20}$$

Multiplicando e dividindo o primeiro termo do lado direito da expressão (4.19) por q_{0i}, obtemos a distribuição condicional para θ_i, dada por

$$(\theta_i|\theta_{-i}, y_i, M, G_0) \sim bMq_{0i}P\,(\theta_i|y_i) + b\sum_{\substack{j=1\\ j\neq i}}^{n} f\,(y_i|\theta_j)\,I_{\theta_j} \tag{4.21}$$

onde

$$P\,(\theta_i|y_i) = \frac{g_0(\theta_i)f(y_i;\theta_i)}{\displaystyle\int g_0(\theta_i)f\,(y_i;\theta_i)\,d(\theta_i)} \propto g_0(\theta_i)f(y_i;\theta_i) \tag{4.22}$$

é a distribuição *a posteriori* de θ_i dado y_i e

$$bMq_{0i} \text{ e } b\sum_{\substack{j=1\\ j\neq i}}^{n} f\,(y_i|\theta_j)$$

são as probabilidades associadas as componentes da mistura em (4.21), com

$$bMq_{0i} + b\sum_{\substack{j=1\\ j\neq i}}^{n} f\,(y_i|\theta_j) = 1.$$

Logo, a constante normalizadora apropriada é dada por

$$b = \left(Mq_{0i} + \sum_{\substack{j=1\\ j\neq i}}^{n} f\,(y_i|\theta_j)\right)^{-1} \tag{4.23}$$

para $i = 1, 2, ..., n$.

Escrevendo (4.21) de forma a mostrar a mistura natural obtida para a distribuição condicional de θ_i, temos

$$(\theta_i|\theta_{-i}, y_i, M, G_0) \begin{cases} = \quad \theta_j & \text{, com probabilidade } bf\,(y_i|\theta_j) \\ \sim P\,(\theta_i|y_i) & \text{, com probabilidade } bMq_{0i} \end{cases} \tag{4.24}$$

para $i = 1, 2, ..., n$ e $j = 1, ..., i-1, i+1, ...n$.

Dependendo da escolha da distribuição base G_0 o modelo MPD é dito ser conjugado ou não conjugado. Em um modelo conjugado as densidades f e g_0 são conjugadas e a integração envolvida no cálculo de q_{0i} pode ser feita analiticamente. Se este não for o caso, o modelo MPD é dito ser não conjugado e as inferências tornam-se mais difíceis, pois a integração envolvida no cálculo de q_{0i}, muitas vezes, é complicada de ser resolvida analiticamente. Porém já estão disponíveis na literatura algoritmos satisfatórios para a resolução deste tipo de problema (ver MacEachern e Müller, 1998 e Neal, 1998).

4.4.1 *Gibbs Sampling* para modelos conjugados

Quando estamos trabalhando com modelos em que exista conjugação entre as densidades f e g_0, todos os cálculos necessários para se obter as distribuições condicionais dos $\theta_{i's}$ podem ser feitos analiticamente e consequentemente a amostragem de θ_i se torna fácil. Logo, podemos utilizar o método *Gibbs Sampling* para se obter amostras das distribuições condicionais de θ_i como a seguir:

- **Algoritmo 1**: Considere uma cadeia de Markov consistindo de $\theta_1, \theta_2, ..., \theta_n$. Amostre repetidamente de (4.24) até atingir convergência.

Caso 1: Variância conhecida

Para ilustrar a aplicação do modelo MPD e do algoritmo 1, considere que as observações $y_1, y_2, ..., y_n$ sejam distribuídas como uma mistura de distribuições normais com médias μ_i e variâncias (conhecidas e constantes) iguais a σ^2, isto é, $y_i \sim N(\mu_i, \sigma^2)$ e a seguinte modelagem utilizando o modelo MPD

$$y_i|\mu_i \sim N(\mu_i, \sigma^2) \tag{4.25}$$

$$\mu_i|G \sim G(\mu_i)$$

$$G|M, G_0 \sim PD\left(Mg_0(\mu_i)\right)$$

com $G_0(\mu_i) = N(0, \tau^2)$ e $g_0(\mu_i) = f_{N(0,\tau^2)}(\mu_i)$ para τ conhecido e $M = c$ (constante).

A densidade condicional para μ_i, como em (4.19), é dada por

$$\left(\mu_i|\mu_{-i}, y_i, M, G_0\right) = bMf_{N(0,\tau^2)}(\mu_i)f_{N(\mu_i,\sigma^2)}(y_i) + b\sum_{\substack{j=1\\j\neq i}}^{n} f_{N(\mu_j,\sigma^2)}(y_i)I_{\mu_j}.$$

Efetuando os cálculos de (4.20) e (4.22), temos que

$$\begin{aligned}
q_{0i} &= \int_{-\infty}^{\infty} f_{N(0,\tau^2)}(\mu_i)f_{N(\mu_i,\sigma^2)}(y_i)d\mu_i \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{1}{2\tau^2}\mu_i^2\right\}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y_i-\mu_i\right)^2\right\}d\mu_i \\
&= \frac{1}{2\pi}\left(\tau^2\sigma^2\right)^{-\frac{1}{2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\tau^2}\mu_i^2 - \frac{1}{2\sigma^2}\left(y_i-\mu_i\right)^2\right\}d\mu_i \\
&= \frac{1}{2\pi}\left(\tau^2\sigma^2\right)^{-\frac{1}{2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{\sigma^2}{\tau^2}\mu_i^2 + \mu_i^2 - 2\mu_i y_i\right)\right\}\exp\left\{-\frac{1}{2\sigma^2}y_i^2\right\}d\mu_i \\
&= \frac{1}{2\pi}\left(\tau^2\sigma^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}y_i^2\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}\left[\left(\frac{\sigma^2+\tau^2}{\tau^2}\right)\mu_i^2 - 2\mu_i y_i\right]\right\}d\mu_i \\
&= \frac{1}{2\pi}\left(\tau^2\sigma^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}y_i^2\right\}\int_{-\infty}^{\infty}\exp\left\{-\frac{\sigma^2+\tau^2}{2\sigma^2\tau^2}\left[\mu_i^2 - 2\mu_i\left(\frac{\tau^2 y_i}{\sigma^2+\tau^2}\right)\right]\right\}d\mu_i \\
&= \frac{1}{2\pi}\left(\tau^2\sigma^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{y_i^2}{2\sigma^2} + \frac{\tau^2 y_i^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right\} \\
&\quad \int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}\left[\mu_i^2 - \left(\frac{\tau^2 y_i}{\sigma^2+\tau^2}\right)\right]^2\right\}d\mu_i \\
&= \frac{1}{2\pi}\left(\tau^2\sigma^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{y_i^2}{2\sigma^2} + \frac{\tau^2 y_i^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right\}\sqrt{\frac{2\pi\sigma^2\tau^2}{\sigma^2+\tau^2}} \\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\tau^2\right)}}\exp\left\{\frac{-\sigma^2 y_i^2 - \tau^2 y_i^2 + \tau^2 y_i^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right\} \\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\tau^2\right)}}\exp\left\{-\frac{1}{2\left(\sigma^2+\tau^2\right)}y_i^2\right\}
\end{aligned}$$

e portanto

$$q_{0i} = f_{N(0,\sigma^2+\tau^2)}(y_i), \tag{4.26}$$

isto é, q_{0i} é o valor da densidade da distribuição normal com média zero e variância $\sigma^2+\tau^2$ no ponto y_i e

$$\begin{aligned} P(\mu_i|y_i) &\propto f_{N(0,\tau^2)}(\mu_i) f_{N(\mu_i,\sigma^2)}(y_i) \\ &\propto \exp\left\{-\frac{1}{2\tau^2}\mu_i^2\right\} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu_i)^2\right\} \end{aligned} \tag{4.27}$$

$$\begin{aligned} &\propto \exp\left\{-\frac{1}{2\tau^2}\mu_i^2 - \frac{1}{2\sigma^2}y_i^2 + \frac{2\mu_i y_i}{2\sigma^2} - \frac{1}{2\sigma^2}\mu_i^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{\sigma^2}{\tau^2}\mu_i^2 + \mu_i^2 - 2\mu_i y_i\right)\right\} \\ &\propto \exp\left\{-\frac{\sigma^2+\tau^2}{2\sigma^2\tau^2}\left[\mu_i^2 - 2\mu_i\left(\frac{\tau^2 y_i}{\sigma^2+\tau^2}\right)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}\left(\mu_i - \frac{\tau^2 y_i}{\sigma^2+\tau^2}\right)^2\right\} \\ &= N\left(\mu_i; \frac{\tau^2 y_i}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) \end{aligned}$$

isto é, a distribuição *a posteriori* de μ_i dado y_i é a distribuição normal com média $\frac{\tau^2 y_i}{\sigma^2+\tau^2}$ e variância $\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$, para $i = 1, 2, ..., n$.

A constante normalizadora apropriada é dada por

$$b = \left(M f_{N(0,\sigma^2+\tau^2)}(y_i) + \sum_{\substack{j=1 \\ j\neq i}}^{n} f_{N(\mu_j,\sigma^2)}(y_i) \right)^{-1},$$

para $i = 1, 2, ..., n$.

A distribuição condicional de μ_i, como em (4.24), é dada por

$$(\mu_i|\mu_{-i}, y_i, M, G_0) \begin{cases} = \mu_j \text{ , com probabilidade } b f_{N(\mu_j,\sigma^2)}(y_i) \\ \sim N\left(\frac{\tau^2 y_i}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right) \text{, com probabilidade } bM f_{N(0,\sigma^2+\tau^2)}(y_i) \end{cases}$$

para $i = 1, 2, ..., n$ e $j = 1, ..., i-1, i+1, ..., n$.

Para o algoritmo 1, inicializamos o método *Gibbs Sampling*, fazendo $\mu_i^0 = y_i$, $i = 1, 2, ..., n$, e a $p-ésima$ seqüência gerada pelo método *Gibbs Sampling* é feita da seguinte maneira:

$$\begin{aligned}
&\text{Amostrar } \mu_1 \text{ de } \left[\mu_1|\mu_2 = \mu_2^{p-1}, \mu_3 = \mu_3^{p-1}, ..., \mu_n = \mu_n^{p-1}\right] \\
&\text{Amostrar } \mu_2 \text{ de } \left[\mu_2|\mu_1 = \mu_1^{p}, \mu_3 = \mu_3^{p-1}, ..., \mu_n = \mu_n^{p-1}\right] \\
&\vdots \\
&\text{Amostrar } \mu_n \text{ de } \left[\mu_n|\mu_1 = \mu_1^{p}, \mu_2 = \mu_2^{p}, ..., \mu_{n-1} = \mu_{n-1}^{p}\right]
\end{aligned} \tag{4.28}$$

Exemplo: Variância conhecida

Para ilustrar numericamente o exemplo acima, desenvolvemos um estudo de simulação utilizando a linguagem R, com o objetivo de verificar a eficiência da modelagem MPD na estimação da média de uma distribuição normal com variância conhecida. Para isto geramos cinco observações de uma distribuição normal com média 2 e variância 1, $N(2, 1)$, e aplicamos a modelagem MPD, como em (4.25), ou seja

$$\begin{aligned}
y_i|\mu_i &\sim N(\mu_i, 1) \\
\mu_i|G &\sim G(\mu_i) \\
G|M, G_0 &\sim PD(Mg_0).
\end{aligned}$$

com $G_0(\mu_i) = N(0, \tau^2)$ e $g_0(\mu_i) = f_{N(0,\tau^2)}(\mu_i)$, para $i = 1, ..., 5$.

Como os cinco valores foram gerados de uma mesma distribuição normal, esperamos que a modelagem MPD detecte estes cinco valores como pertencentes a uma mesma distribuição normal, $N(2, 1)$.

Consideramos $\tau^2 = 25$ de forma a termos pouca informação *a priori* sobre o parâmetro. Devido a isto, escolhemos um valor pequeno para M, pois como temos pouca informação *a priori* sobre os parâmetros devemos ter pouca confiança na distribuição base G_0. Fixamos $M = 1$.

Assim, de (4.26) e (4.27), temos que

$$q_{0i} = f_{N(0,26)}(y_i),\ P(\mu_i|y_i) = N\left(\mu_i; \frac{25}{26}y_i, \frac{25}{26}\right)$$

e

$$b = \left(M f_{N(0,26)}(y_i) + \sum_{\substack{j=1 \\ j \neq i}}^{n} f_{N(\mu_j,1)}(y_i) \right)^{-1}$$

para $i = 1, 2, 3, 4, 5$ e $j \neq i$.

A distribuição condicional de μ_i, como em (4.24), é dada por

$$\left(\mu_i | \mu_{-i}, y_i, M, G_0\right) \begin{cases} = \mu_j & \text{, com probabilidade } b f_{N(\mu_j,1)}(y_i) \\ \sim P(\mu_i | y_i) & \text{, com probabilidade } b M f_{N(0,26)}(y_i) \end{cases}$$

para $i = 1, 2, 3, 4, 5$ e $j \neq i$.

Definidas as distribuições condicionais aplicamos o método *Gibbs Sampling* como descrito em (4.28). Ou seja

$$\begin{aligned}
&\text{Amostramos } \mu_1 \text{ de } \left[\mu_1 | \mu_2 = \mu_2^{p-1}, \mu_3 = \mu_3^{p-1}, \mu_4 = \mu_4^{p-1}, \mu_5 = \mu_5^{p-1}\right] \\
&\text{Amostramos } \mu_2 \text{ de } \left[\mu_2 | \mu_1 = \mu_1^{p}, \mu_3 = \mu_3^{p-1}, \mu_4 = \mu_4^{p-1}, \mu_5 = \mu_5^{p-1}\right] \\
&\text{Amostramos } \mu_3 \text{ de } \left[\mu_3 | \mu_1 = \mu_2^{p}, \mu_2 = \mu_2^{p}, \mu_4 = \mu_4^{p-1}, \mu_5 = \mu_5^{p-1}\right] \\
&\text{Amostramos } \mu_4 \text{ de } \left[\mu_4 | \mu_1 = \mu_1^{p}, \mu_2 = \mu_2^{p}, \mu_3 = \mu_3^{p}, \mu_5 = \mu_5^{p-1}\right] \\
&\text{Amostramos } \mu_5 \text{ de } \left[\mu_5 | \mu_1 = \mu_1^{p}, \mu_2 = \mu_2^{p}, \mu_3 = \mu_3^{p}, \mu_4 = \mu_4^{p}\right]
\end{aligned}$$

onde p é a $p - ésima$ seqüência gerada pelo método *Gibbs Sampling*.

Para a análise de convergência geramos 2 cadeias, cada uma com 30.000 amostras das quais as primeiras 15.000 foram descartadas e consideramos um salto de 5. Com isso, obtemos para cada uma das cadeias uma amostra de tamanho 2.500. Os resultados obtidos para cada um dos parâmetros (média e intervalo de credibilidade) foram baseados nos valores gerados para as duas cadeias, ou seja, a média e o intervalo de credibilidade foram calculados baseados nos 5.000 valores (2.500 da 1^a cadeia + 2.500 da 2^a cadeia). Para verificar a convergência utilizamos o diagnóstico de Gelman e Rubin disponível no CODA (ver, Best *et al.*, 1995). Os resultados obtidos estão apresentados na Tabela 13.

Os valores obtidos para média da distribuição de cada uma das observações é próxima da verdadeira média, da distribuição da qual os valores foram gerados, $N(2, 1)$. O verdadeiro valor da média está contido nos intervalos de credibilidade (95%). Baseados

nos intervalos de credibilidade (95%), podemos concluir que não há evidências para uma diferença significativa entre os valores estimados para os parâmetros, e conseqüentemente que não há evidências de que as cinco observações pertençam a distribuições de probabilidades diferentes, devido a intersecção apresentada pelos intervalos de credibilidade (95%).

A Figura 10 mostra que as densidades marginais são muito parecidas (praticamente se confundem), levando à mesma conclusão. Também podemos notar a semelhança com a distribuição $N(2,1)$ usada na geração dos dados.

A Tabela 14 mostra a proporção de vezes, baseados nos 5.000 valores gerados, que os parâmetros foram considerados iguais e a proporção de vezes que um novo valor foi gerado de sua distribuição *a posteriori*. Como a proporção de vezes que os parâmetros, de cada uma das observações, foram considerados diferentes, isto é, que um novo valor foi gerado de sua distribuição *a posteriori*, $P(\mu_i|y_i)$, é menor que a proporção de vezes que os parâmetros foram considerados iguais, podemos concluir que não há evidências para diferença entre os valores estimados, e consequentemente que não há evidências de que as observações pertençam a distribuições de probabilidades diferentes.

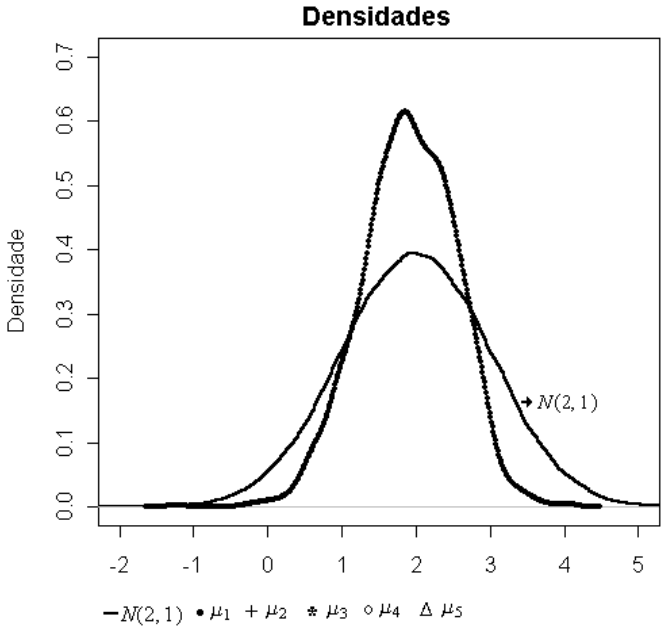

Figura 10: Densidades marginais.

Tabela 13: Valores obtidos das distribuições *a posteriori* condicionais.

Parâmetros	Média	Int. de Cred. (95%)	Diag. Gelmam e Rubin
μ_1	1,9091	(0,5986 ; 3,1063)	1
μ_2	1,9875	(0,7129 ; 3,2542)	1
μ_3	1,9631	(0,6318 ; 3,2415)	1
μ_4	2,0628	(0,7265 ; 3,5896)	1
μ_5	1,9797	(0,6352 ; 3,3775)	1

Tabela 14: Proporção de vezes que os parâmetros foram considerados iguais e proporção de novos valores gerados da distribuição *a posteriori*.

Parâmetros	$= \mu_1$	$= \mu_2$	$= \mu_3$	$= \mu_4$	$= \mu_5$	$= P(\mu_i \vert y_i)$
μ_1	0,0000	0,2379	0,2360	0,2261	0,2347	0,0652
μ_2	0,2381	0,0000	0,2400	0,2341	0,2319	0,0558
μ_3	0,2428	0,2384	0,0000	0,2326	0,2316	0,0546
μ_4	0,2235	0,2351	0,2381	0,0000	0,2280	0,0801
μ_5	0,2367	0,2382	0,2374	0,2337	0,0000	0,0539

Caso 2: Variâncias desconhecidas Para o caso de variâncias desconhecidas, a modelagem MPD, como em (4.16) é dada por

$$\begin{aligned} y_i|\mu_i, \sigma_i^2 &\sim N\left(\mu_i, \sigma_i^2\right) \\ \mu_i, \sigma_i^2|G &\sim G \\ G|M, G_0 &\sim PD\left(MG_0\right) \end{aligned} \tag{4.29}$$

com

$$M = c,\ G_0\left(\mu_i|\sigma_i^2\right) = N\left(\mu_i; 0, \frac{\sigma_i^2}{\lambda}\right),\ G_0\left(\sigma_i^2\right) = IG\left(\sigma_i^2; \frac{\alpha}{2}, \frac{\beta}{2}\right)$$

e portanto

$$G_0\left(\mu_i, \sigma_i^2\right) = N\left(\mu_i; 0, \frac{\sigma_i^2}{\lambda}\right) IG\left(\sigma_i^2; \frac{\alpha}{2}, \frac{\beta}{2}\right)$$

e

$$g_0\left(\mu_i, \sigma_i^2\right) = f_{N\left(0, \frac{\sigma_i^2}{\lambda}\right)}\left(\mu_i\right) f_{IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)}\left(\sigma_i^2\right)$$

para $i = 1, 2, ..., n$.

A densidade condicional de μ_i, σ_i^2, como em (4.19), é dada por

$$\begin{aligned}\left(\mu_i, \sigma_i^2 | \mu_{-i}, \sigma_{-i}^2, y, M, G_0\right) &= bM f_{N\left(0, \frac{\sigma_i^2}{\lambda}\right)}\left(\mu_i\right) f_{IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)}\left(\sigma_i^2\right) f_{N(\mu_i, \sigma_i^2)}\left(y_i\right) \\ &\quad + b \sum_{\substack{j=1 \\ j \neq i}}^{n} f_{N(\mu_j, \sigma_j^2)}\left(y_i\right) I_{\mu_j, \sigma_j^2}.\end{aligned}$$

Efetuando o cálculo de q_{0i}, temos

$$\begin{aligned}q_{0i} &= \int_0^{\infty} f_{N\left(0, \frac{\sigma_i^2}{\lambda}\right)}\left(\mu_i\right) f_{IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)}\left(\sigma_i^2\right) f_{N(\mu_i, \sigma_i^2)}\left(y_i\right) d\mu_i d\sigma_i^2 \\ &= \int_0^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2\sigma_i^2}\left(y_i - \mu_i\right)^2\right\} \frac{1}{\sqrt{2\pi\frac{\sigma_i^2}{\lambda}}} \exp\left\{-\frac{1}{2\frac{\sigma_i^2}{\lambda}}\mu_i^2\right\} \\ &\quad \frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_i^2}\right\} d\mu_i d\sigma_i^2 \\ &= \frac{1}{2\pi}\lambda^{\frac{1}{2}} \frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)} \int_0^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(y_i - \mu_i\right)^2 + \lambda\mu_i^2\right]\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1+1\right)} \\ &\quad \exp\left\{-\frac{\beta}{2\sigma_i^2}\right\} d\mu_i d\sigma_i^2 \\ &= \frac{1}{2\pi}\lambda^{\frac{1}{2}} \frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)} \int_0^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i^2 - 2\mu_i y_i + \mu_i^2 + \lambda\mu_i^2\right)\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1+1\right)} \\ &\quad \exp\left\{-\frac{\beta}{2\sigma_i^2}\right\} d\mu_i d\sigma_i^2 \\ &= \frac{1}{2\pi}\lambda^{\frac{1}{2}} \frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)} \int_0^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\lambda+1\right)\mu_i^2 - 2\mu_i y_i\right]\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1+1\right)} \\ &\quad \exp\left\{-\frac{\beta}{2\sigma_i^2} - \frac{y_i^2}{2\sigma_i^2}\right\} d\mu_i d\sigma_i^2\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\pi}\lambda^{\frac{1}{2}}\frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)}\int_0^{\infty}\int_{-\infty}^{\infty}\exp\left\{-\frac{\lambda+1}{2\sigma^2}\left[\mu_i^2-2\mu_i\left(\frac{y_i}{\lambda+1}\right)\right]\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1+1\right)}\\
&\quad \exp\left\{-\frac{\beta}{2\sigma_i^2}-\frac{y_i^2}{2\sigma_i^2}\right\}d\mu_i d\sigma_i^2\\
&= \frac{1}{2\pi}\lambda^{\frac{1}{2}}\frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)}\int_0^{\infty}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\frac{\sigma^2}{\lambda+1}}\left(\mu_i-\frac{y_i}{\lambda+1}\right)^2\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1+1\right)}\\
&\quad \exp\left\{-\frac{\beta}{2\sigma_i^2}-\frac{y_i^2}{2\sigma_i^2}+\frac{y_i^2}{2\sigma_i^2\left(\lambda+1\right)}\right\}d\mu_i d\sigma_i^2\\
&= \frac{1}{2\pi}\lambda^{\frac{1}{2}}\frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)}\int_0^{\infty}\left(\frac{2\pi\sigma_i^2}{\lambda+1}\right)^{\frac{1}{2}}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1+1\right)}\\
&\quad \exp\left\{-\frac{1}{\sigma_i^2}\left(\frac{\beta\left(\lambda+1\right)+y_i^2(\lambda+1)-y_i^2}{2\left(\lambda+1\right)}\right)\right\}d\sigma_i^2\\
&= \left(\frac{\lambda}{2\pi\left(\lambda+1\right)}\right)^{\frac{1}{2}}\frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)}\int_0^{\infty}\left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+\frac{1}{2}+1\right)}\exp\left\{-\frac{1}{\sigma_i^2}\left(\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right)\right\}d\sigma_i^2\\
&= \left(\frac{\lambda}{2\pi\left(\lambda+1\right)}\right)^{\frac{1}{2}}\frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)}\Gamma\left(\frac{\alpha+1}{2}\right)\left(\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right)^{-\left(\frac{\alpha+1}{2}\right)}\\
&= \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\left(\frac{\lambda}{2\pi\left(\lambda+1\right)}\right)^{\frac{1}{2}}\frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}+\frac{1}{2}}}{\left(\frac{\beta}{2}\right)^{\frac{1}{2}}}\left(\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right)^{-\left(\frac{\alpha+1}{2}\right)}\\
&= \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\left(\frac{\lambda}{\pi\beta\left(\lambda+1\right)}\right)^{\frac{1}{2}}\left(\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{\beta\left(\lambda+1\right)}\right)^{-\left(\frac{\alpha+1}{2}\right)}\\
&= \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\left(\frac{1}{\alpha\pi\frac{\beta(\lambda+1)}{\alpha\lambda}}\right)^{\frac{1}{2}}\left[1+\frac{1}{\alpha}\left(\frac{y_i^2}{\frac{\beta(\lambda+1)}{\alpha\lambda}}\right)\right]^{-\left(\frac{\alpha+1}{2}\right)}\\
&= f_{t_\alpha\left(0,\frac{\beta(\lambda+1)}{\alpha\lambda}\right)}\left(y_i\right)
\end{aligned}
$$

isto é, q_{0i} representa o valor da densidade da distribuição $t-Student$ com média zero e variância $\frac{\beta(\lambda+1)}{\alpha\lambda}$ no ponto y_i com α graus de liberdade, para $i=1,2,...,n$.

Com distribuição conjunta dada por

$$
\begin{aligned}
P\left(\mu_i,\sigma_i^2|y_i\right) &\propto f_{N\left(\mu_i,\sigma_i^2\right)}\left(y_i\right)f_{N\left(0,\frac{\sigma_i^2}{\lambda}\right)}\left(\mu_i\right)f_{IG\left(\frac{\alpha}{2},\frac{\beta}{2}\right)}\left(\sigma_i^2\right)\\
&\propto \left(\sigma_i^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_i^2}\left(y_i-\mu_i\right)^2\right\}\left(\sigma_i^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\frac{\sigma_i^2}{\lambda}}\mu_i^2\right\}\\
&\quad \left(\sigma_i^2\right)^{-\left(\frac{\alpha}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_i^2}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2\sigma_i^2}\left[\left(y_i-\mu_i^2\right)+\lambda\mu_i^2\right]\right\} \\
&\quad \left(\sigma_i^2\right)^{-\left(\frac{\alpha+1+1}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_i^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_i^2}\left(y_i^2-2\mu_i y_i+\mu_i^2+\lambda\mu_i^2\right)\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha+2}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_i^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_i^2}\left[\left(\lambda+1\right)\mu_i^2-2\mu_i y_i\right]\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha+2}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_i^2}-\frac{y_i^2}{2\sigma_i^2}\right\} \\
&\propto \exp\left\{-\frac{\lambda+1}{2\sigma_i^2}\left[\mu_i^2-2\mu_i\left(\frac{y_i}{\lambda+1}\right)+\left(\frac{y_i}{\lambda+1}\right)^2-\left(\frac{y_i}{\lambda+1}\right)^2\right]\right\} \\
&\quad \left(\sigma_i^2\right)^{-\left(\frac{\alpha+2}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_i^2}-\frac{y_i^2}{2\sigma_i^2}\right\} \\
&\propto \exp\left\{-\frac{\lambda+1}{2\sigma_i^2}\left[\mu_i^2-2\mu_i\left(\frac{y_i}{\lambda+1}\right)+\left(\frac{y_i}{\lambda+1}\right)^2\right]\right\} \\
&\quad \left(\sigma_i^2\right)^{-\left(\frac{\alpha+2}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_i^2}-\frac{y_i^2}{2\sigma_i^2}+\frac{y_i^2}{2\sigma_i^2\left(\lambda+1\right)}\right\} \\
&\propto \exp\left\{-\frac{1}{2\frac{\sigma_i^2}{\lambda+1}}\left(\mu_i-\frac{y_i}{\lambda+1}\right)^2\right\}\left(\sigma_i^2\right)^{-\left(\frac{\alpha+2}{2}+1\right)}\exp\left\{-\frac{1}{\sigma_i^2}\left(\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right)\right\} \\
&= N\left(\frac{y_i}{\lambda+1},\frac{\sigma_i^2}{\lambda+1}\right)IG\left(\frac{\alpha+2}{2},\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right)
\end{aligned}
$$

onde $N\left(\frac{y_i}{\lambda+1},\frac{\sigma_i^2}{\lambda+1}\right)$ e $IG\left(\frac{\alpha+2}{2},\frac{\beta(\lambda+1)+\lambda y_i^2}{2(\lambda+1)}\right)$ representam, respectivamente, a distribuição normal com média $\frac{y_i}{\lambda+1}$ e variância $\frac{\sigma_i^2}{\lambda+1}$ e a distribuição gama inversa com parâmetros de forma $\frac{\alpha+2}{2}$ e parâmetro de escala $\frac{\beta(\lambda+1)+\lambda y_i^2}{2(\lambda+1)}$, para $i=1,2,...,n$.

As distribuições condicionais, são dadas por

$$P\left(\mu_i|\sigma_i^2,y_i\right)=N\left(\mu_i;\frac{y_i}{\lambda+1},\frac{\sigma_i^2}{\lambda+1}\right)$$

e

$$P\left(\sigma_i^2|y_i\right)=IG\left(\sigma_i^2;\frac{\alpha+2}{2},\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right)$$

e a constante normalizadora, dada por

$$b=\left(Mf_{t_\alpha\left(0,\frac{\beta(\lambda+1)}{\alpha\lambda}\right)}\left(y_i\right)+\sum_{\substack{j=1\\j\neq i}}^{n}f_{N(\mu_j,\sigma_j^2)}\left(y_i\right)\right)^{-1}$$

para $i = 1, 2, ..., n$ e $j \neq i$.

Logo, como em (4.24), temos que as distribuições condicionais de μ_i, σ_i^2, são dadas por

$$\left(\mu_i, \sigma_i^2 | \mu_{-i}, \sigma_{-i}^2, y_i, M, G_0\right) \begin{cases} = \left(\ \mu_j, \sigma_j^2\right), \text{ com probabilidade } b f_{N(\mu_j, \sigma_j^2)}(y_i) \\ \sim P\left(\mu_i, \sigma_i^2 | y_i\right), \text{ com probabilidade } \ bM f_{t_\alpha\left(0, \frac{\beta(\lambda+1)}{\alpha\lambda}\right)}(y_i) \end{cases} \tag{4.30}$$

com

$$P(\mu_i, \sigma_i^2 | y_i) \ = P(\mu_i | \sigma_i^2, y_i) P(\sigma_i^2 | y_i)$$

onde

$$P(\mu_i | \sigma_i^2, y_i) = N\left(\frac{y_i}{\lambda+1}, \frac{\sigma_i^2}{\lambda+1}\right) \text{ e } P(\sigma_i^2 | y_i) = IG\left(\frac{\alpha+2}{2}, \frac{\beta(\lambda+1) + \lambda y_i^2}{2(\lambda+1)}\right)$$

Definidas as distribuições condicionais para μ_i, σ_i^2, inicializamos o método *Gibbs Sampling* fazendo $\mu_i^0 = y_i$ e $\sigma_i^{2^0} = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$, para $i = 1, 2, ..., n$ e a $p-ésima$ seqüência gerada pelo método *Gibbs Sampling* é feita da seguinte maneira:

$$\begin{array}{lll} \text{Amostrar } \sigma_1^2 \text{ de } \left[\sigma_1^2 | \sigma_2^2 = \sigma_2^{2^{(p-1)}}, \sigma_3^2 = \sigma_3^{2^{(p-1)}}, \cdots, \sigma_n^2 = \sigma_n^{2^{(p-1)}}\right] \\ \text{Amostrar } \sigma_2^2 \text{ de } \left[\sigma_2^2 | \sigma_1^2 = \sigma_1^{2^{(p)}}, \sigma_3^2 = \sigma_3^{2^{(p-1)}}, \cdots, \sigma_n^2 = \sigma_n^{2^{(p-1)}}\right] \\ \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ \text{Amostrar } \sigma_n^2 \text{ de } \left[\sigma_n^2 | \sigma_1^2 = \sigma_1^{2^{(p)}}, \sigma_2^2 = \sigma_2^{2^{(p)}}, \cdots, \sigma_{n-1}^2 = \sigma_{n-1}^{2^{(p)}}\right] \end{array} \tag{4.31}$$

e

$$\begin{array}{lll} \text{Amostrar } \mu_1 \text{ de } \left[\mu_1 | \mu_2 = \mu_2^{p-1}, \mu_3 = \mu_3^{p-1}, \cdots, \mu_n = \mu_n^{p-1} \text{ e } \sigma_1^{2^{(p)}}\right] \\ \text{Amostrar } \mu_2 \text{ de } \left[\mu_2 | \mu_1 = \mu_1^{p}, \mu_3 = \mu_3^{p-1}, \cdots, \mu_n = \mu_n^{p-1} \text{ e } \sigma_2^{2^{(p)}}\right] \\ \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ \text{Amostrar } \mu_n \text{ de } \left[\mu_n | \mu_1 = \mu_1^{p}, \mu_2 = \mu_2^{p}, \cdots, \mu_n = \mu_{n-1}^{p} \text{ e } \sigma_{n-1}^{2^{(p)}}\right]. \end{array} \tag{4.32}$$

Exemplo: Variância desconhecida

Para o procedimento de simulação, consideramos os mesmos valores gerados na simulação descrita acima, isto é, $y_i \sim N(2, 1)$, para $i = 1, 2, 3, 4, 5$, e utilizamos a modelagem

como em (4.29), com $M = 1$ e os hiperparâmetros da distribuição base G_0 foram escolhidos de forma a termos uma distribuição não informativa, para isto fizemos $\lambda = 0,01$; $\alpha = 2$ e $\beta = 2$.

As distribuições condicionais para μ_i, σ_i^2, como em (4.30), são dadas por

$$\left(\mu_i, \sigma_i^2 | \mu_{-i}, \sigma^2_{-i}, , y_i, M, G_0\right) \begin{cases} = \quad \left(\mu_j, \sigma_j^2\right), \text{ com probabilidade } bf_{N(\mu_j, \sigma_j^2)}(y_i) \\ \sim P(\mu_i, \sigma_i^2 | y_i) \quad , \text{ com probabilidade } bMf_{t_\alpha(0,101)}(y_i) \end{cases}$$

com

$$P(\mu_i, \sigma_i^2 | y_i) \ = P(\mu_i | \sigma_i^2, y_i) P(\sigma_i^2 | y_i)$$

onde

$$P(\mu_i | \sigma_i^2, y_i) = N\left(\frac{y_i}{1,01}, \frac{\sigma_i^2}{1,01}\right) \text{ e } P(\sigma_i^2 | y_i) = IG\left(2,1 + \frac{0,01}{2,02} {y_i}^2\right)$$

Definidas as distribuições condicionais, aplicamos o método *Gibbs Sampling* como descrito em (4.31) e (4.32).

Para análise de convergência geramos 2 cadeias, cada uma com 70.000 amostra. Devido a não obtenção de convergência quando utilizamos 2 cadeias com 30.000 amostras, descartamos as primeiras 35.000 e consideramos um salto de 10. Os resultados apresentados nas Tabelas 15 e 16 foram obtidos a partir dos valores gerados para as duas cadeias. Para verificação de convergência, utilizamos o diagnóstico de Gelmam e Rubin disponível no recurso CODA.

Os valores obtidos para a média e variância de cada uma das observações é pró- xima da verdadeira média e variância, da distribuição da qual os valores foram gerados. O verdadeiro valor da média e da variância está contido nos intervalos de credibilidade (95%). Baseados nos intervalos de credibilidade (95%), na Figura 11 e nas proporções apresentadas na Tabela 17, podemos concluir pela não evidência para uma diferença significativa entre os valores estimados para as médias e variâncias das observações, devido a intersecção que os intervalos apresentam, pela semelhança das densidades marginais e pela proporção de vezes que os parâmetros foram considerados iguais. Conseqüentemente concluímos pela não evidência de que as observações pertençam a distribuições de

probabilidades diferentes.

Tabela 15: Valores médios das distribuições *a posteriori*.

Parâmetros	Média	Int. de Cred. (95%)	Diag. Gelmam e Rubin
σ_1^2	0,9696	(0,2002 ; 3,7043)	1
σ_2^2	0,9685	(0,1977 ; 3,6530)	1
σ_3^2	1,0161	(0,2002 ; 3,9172)	1
σ_4^2	1,0526	(0,2047 ; 4,0295)	1
σ_5^2	1,036	(0,2002 ; 4,0414)	1

Tabela 16: Valores médios das distribuições *a posteriori*.

Parâmetros	Média	Int. de Cred. (95%)	Diag. Gelmam e Rubin
μ_1	1,8632	(0,4388 ; 3,0301)	1
μ_2	1,9408	(0,6189 ; 3,1721)	1
μ_3	1,9351	(0,5820 ; 3,2332)	1
μ_4	2,0027	(0,5271 ; 3,6268)	1
μ_5	1,9531	(0,5203 ; 3,3625)	1

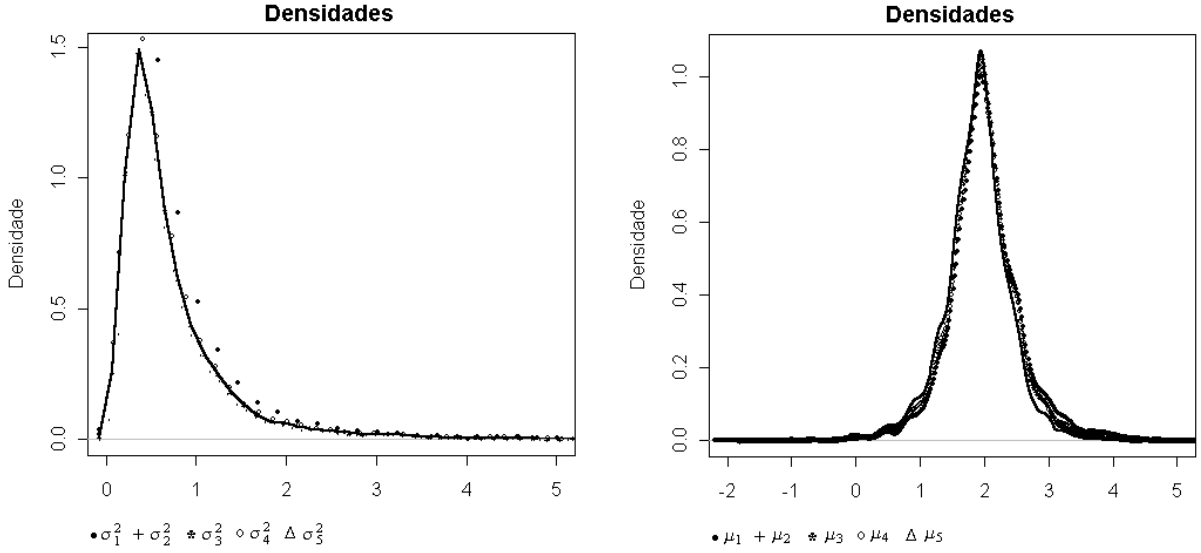

Figura 11: Densidades marginais *a posteriori* para σ_i^2 e μ_i, $i = 1, 2, 3, 4, 5$.

Tabela 17: Proporção de vezes que os parâmetros foram considerados iguais e proporção de vezes que novos valores foram gerados da distribuição *a posteriori*.

Parâmetros	$=(\mu_1,\sigma_1^2)$	$=(\mu_2,\sigma_2^2)$	$=(\mu_3,\sigma_3^2)$	$=(\mu_4,\sigma_4^2)$	$=(\mu_5,\sigma_5^2)$	$=P(\theta_i\|y_i)$
(μ_1,σ_1^2)	0,0000	0,2307	0,2319	0,2493	0,2296	0,0885
(μ_2,σ_2^2)	0,2410	0,0000	0,2430	0,2363	0,2377	0,0419
(μ_3,σ_3^2)	0,2407	0,2424	0,0000	0,2301	0,2384	0,0483
(μ_4,σ_4^2)	0,2229	0,2443	0,2359	0,0000	0,2391	0,0578
(μ_5,σ_5^2)	0,2386	0,2412	0,2417	0,2361	0,0000	0,0425

4.5 Aplicação: Análise da expressão gênica

Uma vantagem que a abordagem utilizando o modelo de misturas de processos Dirichlet apresenta, para análise da expressão gênica, é que podemos comparar simultaneamente com a situação de controle, mais de uma condição de tratamento, ao invés de apenas uma, como na abordagem utilizando o teste t, o fator de Bayes e o critério DIC, como será mostrado a seguir.

De modo a facilitar o desenvolvimento das análises, introduzimos a seguinte notação.

Considere que para cada gene g, temos um conjunto de variáveis observáveis $x_1^{t_0},...,x_{n_0}^{t_0}$, $x_1^{t_1},...,x_{n_1}^{t_1}$, $x_1^{t_2},...,x_{n_2}^{t_2}$, ..., $x_1^{t_K},...,x_{n_K}^{t_K}$, independentes, representando o logaritmo dos níveis de expressão do gene na situação de controle (t_0), tratamento 1 (t_1) e sucessivamente até a situação de tratamento K (t_K). Ou seja,

- $x_1^{t_0},...,x_{n_0}^{t_0}$: representa as n_0 medidas dos níveis de expressão para o g-ésimo gene sob a condição de controle (t_0);

- $x_1^{t_1},...,x_{n_1}^{t_1}$: representa as n_1 medidas dos níveis de expressão para o g-ésimo gene sob a condição de tratamento 1 (t_1);

$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$

- $x_1^{t_K},...,x_{n_K}^{t_K}$: representa as n_K medidas dos níveis de expressão para o g-ésimo gene sob a condição de tratamento K (t_K).

Como mencionado anteriormente, consideramos que o logaritmo das medidas dos níveis de expressão observados para um gene g, em cada uma das situações t_k, para

$k = 0, 1, ..., K$, como sendo gerados de uma distribuição normal. Logo,

$$\begin{aligned} x_1^{t_0}, ..., x_{n_0}^{t_0} &\sim N(\mu_{t_0}, \sigma_{t_0}^2) \\ x_1^{t_1}, ..., x_{n_1}^{t_1} &\sim N(\mu_{t_1}, \sigma_{t_1}^2) \\ &\vdots \\ x_1^{t_K}, ..., x_{n_K}^{t_K} &\sim N(\mu_{t_K}, \sigma_{t_K}^2) \end{aligned}$$

com os genes sendo tratados independentemente.

Considere $\bar{x}_{gt_0}$, $\bar{x}_{gt_1}, ..., \bar{x}_{gt_K}$, para $g = 1, 2, ..., G$, onde

- $\bar{x}_{gt_0}$: média das medidas dos níveis de expressão observados para o g-ésimo gene na situação de controle (t_0),

- $\bar{x}_{gt_1}$: média das medidas dos níveis de expressão observados para o g-ésimo gene na situação de tratamento 1 (t_1),

$\vdots$ $\vdots$ $\vdots$

- $\bar{x}_{gt_K}$: média das medidas dos níveis de expressão observados para o g-ésimo gene na situação de tratamento K (t_K).

Logo,

$$\begin{aligned} \bar{x}_{gt_0} &\sim N\left(\mu_{t_0}, \frac{\sigma_{t_0}^2}{n_0}\right) \\ \bar{x}_{gt_1} &\sim N\left(\mu_{t_1}, \frac{\sigma_{t_1}^2}{n_1}\right) \\ &\vdots \\ \bar{x}_{gt_K} &\sim N\left(\mu_{t_K}, \frac{\sigma_{t_K}^2}{n_K}\right) \end{aligned}$$

isto é, como conseqüência da suposição de que as variáveis observáveis, $x_1^{t_k}, ..., x_{n_k}^{t_k}$, são geradas segundo uma distribuição normal com média μ_{t_k} e variância $\sigma_{t_k}^2$, temos que suas respectivas médias amostrais são geradas segundo uma distribuição normal com mesma média populacional e variância $\frac{\sigma_{t_k}^2}{n_k}$, para $k = 0, 1, 2, ..., K$.

Considerar se há ou não evidências para níveis de expressão diferentes para um determinado gene g, equivale a considerar se a média observada, $\bar{x}_{gt_k}$ na condição de tratamento t_k, para $k = 1, 2, ..., K$, é gerada ou não da mesma distribuição de probabilidades,

$N\left(\mu_{t_0}, \frac{\sigma^2_{t_k}}{n_k}\right)$ da média $\bar{x}_{gt_0}$ das observações de controle t_0.

Para isso, como em (4.29), assumimos que

$$\begin{aligned} \bar{x}_{gt_k}|\mu_{gt_k}, \sigma^2_{gt_k} &\sim N\left(\mu_{gt_k}, \frac{\sigma^2_{gt_k}}{n_k}\right) \\ \mu_{gt_k}, \sigma^2_{gt_k}|G &\sim G \\ G|M, G_0 &\sim PD\left(MG_0\right) \end{aligned} \tag{4.33}$$

com

$$M = c,\ G_0(\mu_{gt_k}|\sigma^2_{gt_k}) = N\left(0, \frac{\sigma^2_{gt_k}}{n_k\lambda}\right),\ G_0(\sigma^2_{gt_k}) = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

e portanto

$$G_0(\mu_{gt_k}, \sigma^2_{gt_k}) = N\left(0, \frac{\sigma^2_{gt_k}}{n_k\lambda}\right) IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

e

$$g_0(\mu_{gt_k}, \sigma^2_{gt_k}) = f_{N\left(0, \frac{\sigma^2_{gt_k}}{n_k\lambda}\right)}\left(\mu_{gt_k}\right) f_{IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)}\left(\sigma^2_{gt_k}\right).$$

As distribuições condicionais de $\mu_{gt_k}, \sigma^2_{gt_k}$, como em (4.30), são dadas por:

$$\left(\mu_{gt_k}, \sigma^2_{gt_k}|\mu_{-gt_k}, \sigma^2_{-gt_k}, \bar{x}_{gt_k}, M, G_0\right) \begin{cases} = \left(\mu_{gt_j}, \sigma^2_{gt_j}\right), \text{ com prob. } bf_{N\left(\mu_{gt_j}, \frac{\sigma^2_{gt_j}}{n_j}\right)}\left(\bar{x}_{gt_k}\right) \\ \sim P\left(\mu_{gt_k}, \sigma^2_{gt_k}|y_{gt_k}\right), \text{ com prob. } bMf_{t_\alpha\left(0, \frac{\beta(\lambda+1)}{\alpha\lambda n_k}\right)}\left(\bar{x}_{gt_k}\right) \end{cases} \tag{4.34}$$

com

$$P\left(\mu_{gt_k}, \sigma^2_{gt_k}|\bar{x}_{gt_k}\right) = P\left(\mu_{gt_k}|\sigma^2_{gt_k}, \bar{x}_{gt_k}\right) P\left(\sigma^2_{gt_k}|\bar{x}_{gt_k}\right)$$

onde

$$\begin{aligned} P\left(\mu_{gt_k}|\sigma^2_{gt_k}, \bar{x}_{gt_k}\right) &= N\left(\frac{\bar{x}_{gt_k}}{\lambda+1}, \frac{\sigma^2_{gt_k}}{n_k\left(\lambda+1\right)}\right), \\ P\left(\sigma^2_{gt_k}|\bar{x}_{gt_k}\right) &= IG\left(\frac{\alpha+2}{2}, \frac{\beta\left(\lambda+1\right) + n_k\lambda\bar{x}^2_{gt_k}}{2\left(\lambda+1\right)}\right) \end{aligned}$$

e

$$b = \left(Mf_{t_\alpha\left(0, \frac{\beta(\lambda+1)}{\alpha\lambda n_k}\right)}\left(\bar{x}_{gt_k}\right) + \sum_{\substack{j=1 \\ j\neq k}}^{K} f_{N\left(\mu_{gt_j}, \frac{\sigma^2_{gt_j}}{n_j}\right)}\left(\bar{x}_{gt_k}\right)\right)^{-1}$$

para $k = 0, 1, 2, ..., K$ e $j \neq k$.

Para fazermos inferências sobre os possíveis genes que apresentam ou não evidências para diferença, utilizamos o método *Gibbs Sampling*, como descrito em (4.31) e (4.32) e nos baseamos na proporção de vezes que o modelo MPD considerou $\bar{x}_{gt_k}$ e $\bar{x}_{gt_0}$ como sendo gerados ou não de uma mesma distribuição normal. Isto é nos baseamos na proporção de vezes que o modelo MPD considerou $\left(\mu_{t_k}, \sigma^2_{t_k}\right) = \left(\mu_{t_0}, \sigma^2_{t_0}\right)$, $k = 1, 2, ..., K$. Se esta proporção for maior que a proporção de vezes que um novo valor foi gerado da distribuição *a posteriori*, $P\left(\mu_{gt_k}, \sigma^2_{gt_k} | \bar{x}_{gt_k}\right)$, então decidiremos pela não evidência para diferença, caso contrário decidiremos pela evidência para diferença.

4.5.1 Simulação

Desenvolvemos para o modelo MPD, o mesmo estudo de simulação realizado para o teste t, para o fator de Bayes e para o critério DIC, com o objetivo de verificar seu comportamento na detecção de genes com evidências para níveis de expressão diferentes, quando consideramos diferentes afastamentos na média e/ou na variância das observações de tratamento com relação às observações de controle.

Os hiperparâmetros utilizados foram selecionados de forma a obtermos distribuições *a priori* pouco informativas. Para isto, utilizamos $M = 1$; $\lambda = 0,01$; $\alpha = 2$ e $\beta = 2$; $\beta = 1$ e $\beta = 0,5$.

Os diferentes valores para β foram utilizados para verificarmos a sensibilidade do modelo MPD na detecção dos genes com evidências para diferença com relação a escolha dos hiperparâmetros. Para cada valor de β, temos as distribuições condicionais dadas como em (4.34).

Para aplicação do método *Gibbs Sampling*, geramos 10000 amostras das distribuições condicionais (4.34), para cada um dos 1000 genes (simulados) e cada valor de β e calculamos as proporções de vezes que os parâmetros $\left(\mu_{t_k}, \sigma^2_{t_k}\right)$ e $\left(\mu_{t_0}, \sigma^2_{t_0}\right)$ foram considerados iguais e proporção de vezes que novos valores foram gerados da distribuição *a posteriori*, $P\left(\mu_{gt_k}, \sigma^2_{gt_k} | \bar{x}_{gt_k}\right)$, e decidimos pela presença de evidências para diferença ou não, para $k = 1, 2, ...K$.

As Tabelas 18, 19 e 20 mostram as quantidades de genes (simulados) detectados com evidências para níveis de expressão diferentes pelo modelo MPD, respectivamente para

$\beta = 2$; $\beta = 1$ e $\beta = 0,5$, para cada variação δ e γ (ou ambos) considerados. Cada linha mostra as quantidades de genes (simulados) detectados com evidências para diferença com γ fixo e δ variando. Cada coluna mostra as quantidades de genes (simulados) detectados com evidências para diferença com δ fixo e γ variando.

Para γ fixo, cada linha das Tabelas 18, 19 e 20, e a partir de $\delta = 0$, aumentando ou diminuindo o valor de δ, o modelo MPD, para os valores de β utilizados, detecta uma quantidade maior de genes com evidências para níveis de expressão diferentes do que a variação δ anterior. Ou seja, o modelo MPD é sensível a variação na média, detectando os genes com evidências para níveis de expressão diferentes.

Para $\delta = \pm 0,8$ ou $\delta = \pm 1$ fixo, cada coluna das Tabelas 18 e 19, aumentando o valor de γ, o modelo MPD detecta uma quantidade menor de genes com evidências para níveis de expressão diferentes do que a variação γ anterior. O mesmo acontecendo na Tabela 20 ($\beta = 0,5$), com exceção de $\delta = 0$. Esta característica do modelo MPD já era esperada, pois a modelagem (4.33) utilizada, leva em consideração apenas a modelagem das médias de controle e de tratamento.

A Tabela 21 mostra uma estatística descritiva de 10.000 valores gerados (utilizando a linguagem R) das distribuições *a priori* para a variância, $IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$ com $\alpha = 2$ e $\beta = 2$, $\beta = 1$ e $\beta = 0,5$. Dos 10.000 valores gerados com $\beta = 2$, a média dos valores é 11,599 e a variância 48064,5. Para $\beta = 1$ a média dos valores é 5,799 e a variância 12016,1 e para $\beta = 0,5$ a média dos valores é 2,899 e a variância 3004,4. Ou seja, as três distribuição *a priori* para a variância estão sendo pouco informativas, pois as variâncias dos valores gerados são todas "altas".

A medida que diminuímos o valor de β, isto é, temos informação *a priori*, o modelo MPD detecta uma quantidade maior de genes com evidências para diferença (ver Tabelas 19, 19 e 20). Isto justifica o fato de os valores das colunas (ou linha) da Tabela 18 serem menores que os respectivos valores da Tabela 19, que é menor que os respectivos valores da Tabela 20.

Assim, temos que para distribuições *a priori* não informativas, tais como $IG(1;1)$, o modelo MPD pode ser influenciado a não detectar genes com evidências para diferença, quando há uma diferença presente na média e/ou variância das medidas de tratamento com relação as medidas de controle. Logo, mesmo para distribuições *a priori* pouco

informativas, devemos ter certa atenção para definirmos os valores dos hiperparâmetros para obtermos resultados satisfatórios. Se possível devemos nos basear na opinião de um especialista da área genética para definirmos estes valores.

A sensibilidade do modelo MPD, com relação aos hiperparâmetros, é justificada pelo fato do algoritmo *Gibbs Sampling* descrito "depender" da geração de um melhor valor para substituir o presente valor da cadeia de *Markov*. Isto é, se um valor melhor demora para ser gerado, o algoritmo terá uma alta taxa de aceitação e automaticamente uma baixa taxa de geração de novos valores das distribuições *a posteriori* e consequentemente o gene será detectado sem evidência para diferença.

Logo, se temos informação (obtida da opinião de um especialista ou através de métodos empíricos) para definirmos os valores dos hiperparâmetros, temos uma amostragem mais eficiente e melhores resultados com relação à detecção de genes com evidências para diferença.

Assim, acreditamos que a utilização do modelo MPD, comparado ao teste t, fator de Bayes e critério DIC, forneça um acréscimo de qualidade na detecção dos genes com evidências para diferença, com relação aos detectados pelo teste t e que a diferença de seleção de genes entre o modelo MPD e o fator de Bayes e o critério DIC, está na detecção de evidências para níveis de expressão diferentes com relação à variação na variância das medidas de tratamento com relação as medidas de controle, presente nas aplicações do fator de Bayes e do critério DIC.

Tabela 18: Quantidades detectadas com evidência para diferença, com $\beta = 2$.

	δ						
γ	-1,0	-0,8	-0,5	0,0	0,5	0,8	1,0
0,25	998	898	162	0	196	918	998
1	994	861	195	0	226	904	994
4	952	783	275	03	292	800	970
9	880	710	340	25	353	737	913
16	821	668	382	93	384	682	854

Tabela 19: Quantidades detectadas com evidência para diferença, com $\beta = 1$.

	δ						
γ	-1,0	-0,8	-0,5	0,0	0,5	0,8	1,0
0,25	1000	990	424	0	437	997	1000
1	1000	975	437	0	447	981	1000
4	991	891	469	08	469	925	994
9	946	817	471	66	478	849	956
16	895	761	478	161	479	783	914

Tabela 20: Quantidades detectadas com evidência para diferença, com $\beta = 0,5$.

	δ						
γ	-1,0	-0,8	-0,5	0,0	0,5	0,8	1,0
0,25	1000	1000	744	0	771	1000	1000
1	1000	997	710	01	727	999	1000
4	999	965	655	34	641	974	998
9	979	899	607	140	591	912	983
16	932	827	586	256	583	858	941

Tabela 21: Estatística descritiva dos valores gerados da distribuições *a priori*.

Valor β	D. *a priori*	Min	Média	Var	Q. 0,25	Med.	Q. 0,75	Max
$\beta = 2$	$IG(\,(1;1)$	0,112	11,599	48064,5	0,713	1,448	3,482	15472,9
$\beta = 1$	$IG\,(1;0,5)$	0,056	5,799	12016,1	0,357	0,724	1,741	7736,4
$\beta = 0,5$	$IG\,(1;0,25)$	0,028	2,899	3004,4	0,178	0,3620	0,871	3868,3

4.5.2 Aplicação

Aplicamos a modelagem (4.33), no experimento realizado com as células da bactéria *Escherichia Coli* com relação aos padrões IHF^+ e IHF^-, já utilizado na aplicação do teste t, do fator de Bayes e do critério DIC.

Como citado anteriormente, dispomos de 434 genes e uma condição de tratamento. Ou seja, $g = 1, 2, ..., 434$ e $k = 0, 1$. Para cada um dos genes temos cinco medidas de níveis de expressão para a situação de controle (t_0) e cinco medidas de níveis de expressão

para a situação de tratamento (t_1). Isto é,

- $x_1^{t_0}, ..., x_5^{t_0}$: representa as 5 medidas dos níveis de expressão para o g-ésimo gene sob a situação de controle (t_0),

- $x_1^{t_1}, ..., x_5^{t_1}$: representa as 5 medidas dos níveis de expressão para o g-ésimo gene sob a situação de tratamento (t_1),

- $\bar{x}_{gt_0}$: média dos níveis de expressão do g-ésimo gene na situação de controle (t_0),

- $\bar{x}_{gt_1}$: média dos níveis de expressão do g-ésimo gene na situação de tratamento (t_1).

Os hiperparâmetros utilizados, foram os mesmos utilizados na simulação. Para cada valor de β, temos as distribuições condicionais, dadas como em (4.34).

Definidas as distribuições condicionais, aplicamos o método *Gibbs Sampling* como em (4.31) e (4.32). Para cada um dos genes g, $g = 1, 2, ..., 434$, e cada valor de β, geramos 30000 amostras das distribuições condicionais (4.34). Não analisamos convergência pois não estamos interessados na estimação dos parâmetros $\left(\mu_{gt_k}, \sigma^2_{gt_k}\right)$ das medidas dos genes g, mas sim na proporção de vezes que o modelo MPD considerou $\bar{x}_{gt_1}$ e $\bar{x}_{gt_0}$ como sendo gerados de uma mesma distribuição normal durante as $p = 30000$ seqüências geradas pelo método *Gibbs Sampling*. Isto é, estamos interessados na proporção de vezes que o modelo MPD considerou $\left(\mu_{t_k}, \sigma^2_{t_k}\right) = \left(\mu_{t_0}, \sigma^2_{t_0}\right)$ durante as $p = 30000$ seqüências geradas pelo método *Gibbs Sampling*,

As Figuras 12, 13 e 14 mostram as médias controle e tratamento para cada um dos genes g, $g = 1, 2, ..., 434$, destacando os genes g detectados com evidência para níveis de expressão diferentes. Nestas Figuras os pontos $\bullet$ indicam os genes que não foram detectados com evidências para diferença e os sinais + indicam os genes que foram detectados com evidências para diferença.

Para $\beta = 2$; distribuição *a priori* $IG\,(1, 1)$ para a variância, 5 genes foram detectados com evidências para níveis de expressão diferentes (Figura 12).

Para $\beta = 1$; distribuição *a priori* $IG(1; 0, 5)$ para a variância, 13 genes foram detectados com evidência para níveis de expressão diferentes (Figura 13).

Para $\beta = 0, 5$, distribuição *a priori* $IG(1; 0, 25)$ para a variância, 23 genes foram detectados com evidência para níveis de expressão diferentes (Figura 14).

Como as variâncias das medidas dos níveis de expressão dos genes das células da bactéria *Escherichia Coli* são todas altamente concentradas no intervalo (0;0,3), podemos

verificar pelos valores da Tabela 21 que todas as três distribuições *a priori* utilizadas são pouco informativas, o que também pode ser observado pela Figura 15 que mostra as densidades empíricas das variâncias controle e tratamento e as densidades das distribuições *a priori* utilizadas. Para a distribuição *a priori* para a variância $IG(1,1)$, somente 325 (3,25%) dos valores gerados são menores que 0,3. Ou seja, a utilização desta distribuição *a priori* está sendo muito pouco informativa para a utilização, pois poucos valores gerados pertencem ao domínio dos valores observados para a variância. Para os outros valores de β, $\beta = 1$ e $\beta = 0,5$, 19,15% e 43,59%, respectivamente, dos valores gerados são menores que 0,3. Isto é, os valores obtidos com estas duas distribuições *a priori*, $IG(1;0,5)$ e $IG(1;0,25)$, são mais indicados para aplicação aos dados da bactéria, o que reflete nos resultados obtidos, como pode ser observado nas Figuras 13 e 14.

Como citado na simulação, o modelo MPD é sensível a escolha dos hiperparâmetros das distribuições *a priori*, levando a diferentes conclusões, ou seja, a escolha de diferentes genes g com evidências para níveis de expressão diferentes. Esta sensibilidade é justificada pelo fato do algoritmo *Gibbs Sampling* descrito para o modelo MPD "depender"da geração de um melhor valor para substituir o presente valor da cadeia de *Markov*. Dessa forma, mesmo para distribuições *a priori* não informativas devemos ter certa atenção para definirmos os valores dos hiperparâmetros para obtermos resultados satisfatórios. Para definir os valores dos hiperparâmetros de forma adequada podemos nos basear na opinião de um especialista ou utilizar métodos empíricos (ver, Efron *et al.*, 2001 ou Dahl, 2002).

Para distribuições *a priori* não informativas definidas de forma coerente, tais como $IG(1;0,5)$ ou $IG\,(1;0,25)$ para a variância, o modelo MPD detecta evidências para diferença quando a variação está presente na média ou presente na média e na variância das medidas de tratamento com relação às medidas de controle. Um exemplo, é o gene 277, que como já mencionado anteriormente possui uma diferença presente tanto na média quanto na variância, das medidas de tratamento com relação às medidas de controle. Porém quando a diferença está presente somente na variância, das medidas de tratamento com relação as medidas de controle, o modelo MPD não detecta a evidência. Um exemplo é o gene 323 que apresenta diferença apenas com relação às variâncias (ver Figura 15). Isto justifica o fato de o modelo MPD ter detectado apenas os genes com média de tratamento e de controle distante da reta $y = x$ (ver Figuras 13, 14 e 15).

A medida que temos informação *a priori* (diminuímos o valor de β) o modelo MPD detecta uma quantidade maior de genes com evidências para diferença. Porém de uma forma coerente, detectando apenas os genes com média de controle e de tratamento mais distantes da reta $y = x$.

Comparando os resultados obtidos com a aplicação do modelo MPD com os obtidos com a aplicação do teste t, do fator de Bayes e do critério DIC, temos que:

- Os genes 84, 134, 179, 248, 360, 376 e 383 foram detectados pelo modelo MPD e pelo teste t (ver Tabelas 2 e 22);

- Dos 23 genes detectados pelo modelo MPD, os genes 84, 223, e 360 não foram detectados pelo fator de Bayes e os genes 84, 223, 332, 360 e 376 não foram detectados pelo critério DIC (ver Tabelas 8, 12 e 22).

- Os genes 134, 179, 248 e 383 foram detectados pelos quatro métodos (ver Tabelas 2, 8, 12 e 22);

Assim, se temos informação para definirmos os valores dos hiperparâmetros, acreditamos que a utilização do modelo MPD, ofereça um acréscimo de qualidade na detecção de genes com evidências para diferença com relação a utilização do teste t descrito no Capítulo 2. Pois os genes com média de controle e de tratamento mais distantes da reta $y = x$ foram detectados com evidências para diferença, o que não acontece com a aplicação do teste t.

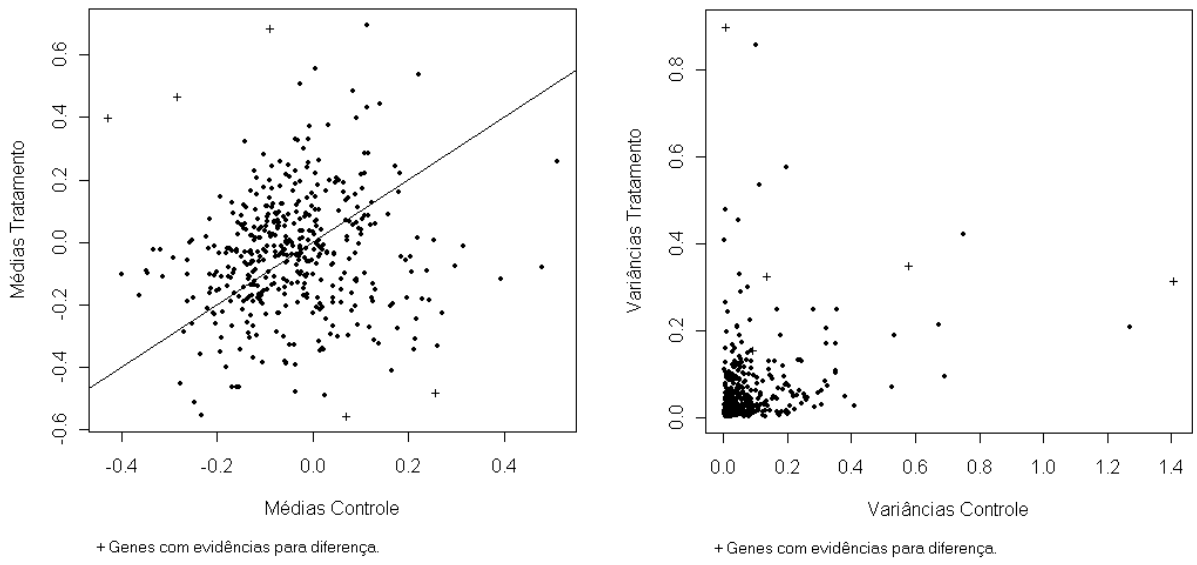

Figura 12: Médias e variâncias controle e tratamento, com $\beta = 2$.

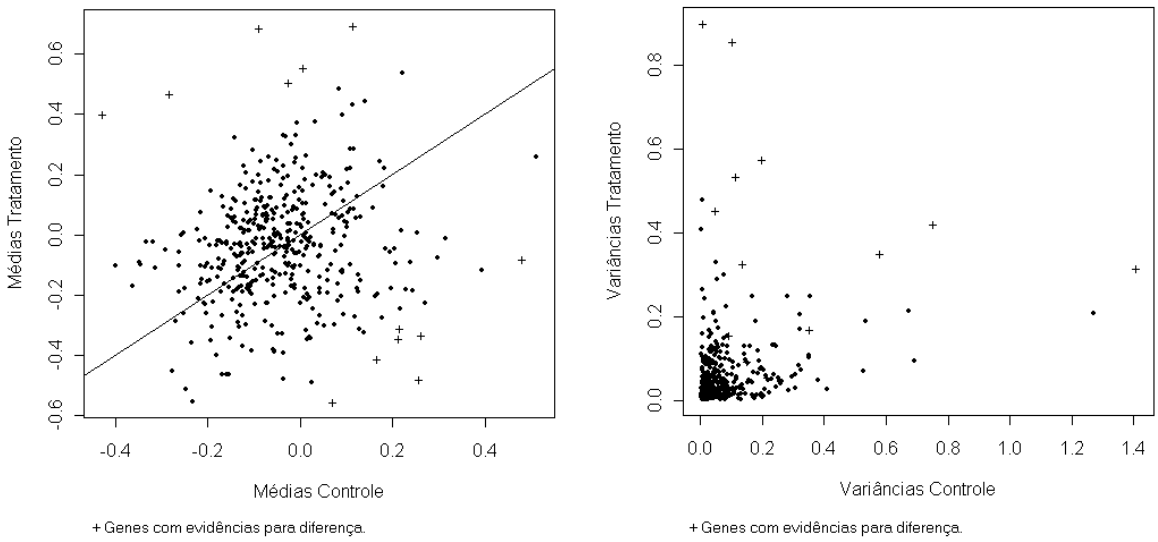

Figura 13: Médias e variâncias controle e tratamento, com $\beta = 1$.

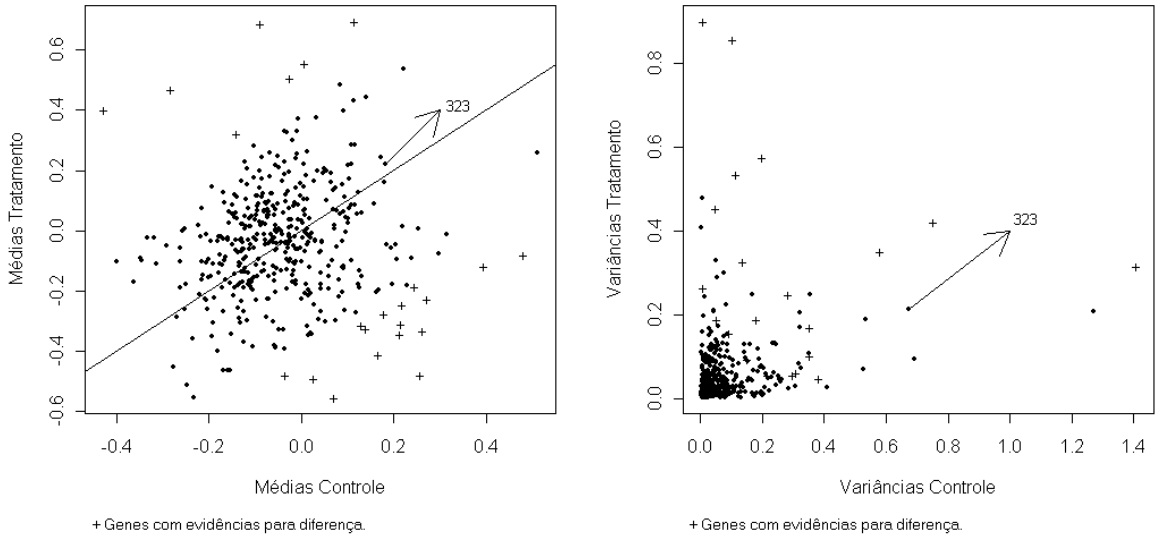

Figura 14: Médias e variâncias controle e tratamento, com $\beta = 0,5$.

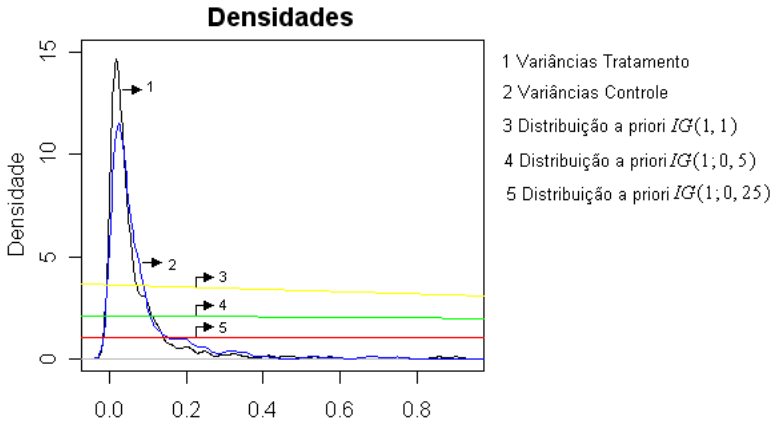

Figura 15: Densidades das variâncias controle e tratamentos e densidades *a priori*.

A Tabela 22, mostra as médias e variâncias controle e tratamento e a numeração dos 23 genes detectados com evidências para diferença nos níveis de expressão quando utilizamos $\beta = 0,5$. Os genes indicados com * são os detectados quando utilizamos $\beta = 1$ e os indicados com $\circ$ são os detectados quando $\beta = 2$.

Tabela 22: Genes detectados com evidências para níveis de expressão diferentes.

Gene	Média controle	Var controle	Média tratamento	Var tratamento
83	0,1393	0,3084	-0,3253	0,0628
84	-0,0359	0,0792	-0,4785	0,0473
*121	0,0064	0,1121	0,5560	0,5337
133	0,3939	0,3811	-0,1173	0,0470
*134	0,2603	0,2152	-0,3331	0,0567
156	0,1278	0,2814	-0,3109	0,2470
$^{*}_{\circ}$179	0,0705	0,1365	-0,5547	0,3277
193	0,1781	0,2976	-0,2767	0,0573
*201	0,1650	0,3507	-0,4115	0,1697
223	-0,0577	0,0563	-0,3796	0,1734
*247	0,4787	0,1960	-0,0782	0,5741
*248	0,2133	0,1489	-0,3095	0,0939
$^{*}_{\circ}$251	-0,4292	1,4075	0,4018	0,3153
*256	-0,0271	0,0454	0,5082	0,4225
*277	0,1144	0,1010	0,6958	0,8571
$^{*}_{\circ}$324	-0,0886	0,0066	0,6860	0,9007
332	-0,1419	0,0071	0,3220	0,2635
360	0,2449	0,1277	-0,1841	0,0318
*256	-0,0271	0,0454	0,5082	0,4225
*277	0,1144	0,1010	0,6958	0,8571
$^{*}_{\circ}$324	-0,0886	0,0066	0,6860	0,9007
332	-0,1419	0,0071	0,3220	0,2635
360	0,2449	0,1277	-0,1841	0,0318
376	0,2174	0,0490	-0,2458	0,1899

Gene	Média controle	Var controle	Média tratamento	Var tratamento
*˚383	0,2599	0,0903	-0,4787	0,1561
*385	0,2114	0,7492	-0,3441	0,4202
415	0,2706	0,3514	-0,2245	0,1032
*˚417	-0,2840	0,5785	0,4681	0,3519

Capítulo 5

Modelo com mistura infinita

Os modelos bayesianos, baseados em modelos com mistura de distribuições formam uma classe de modelos estatísticos que são aplicados quando observações podem vir de mais de uma distribuição de probabilidades. Este tipo de modelagem está sendo utilizado em diversas áreas, especialmente devido à formulação de novos algoritmos computacionais e pela forma natural de interpretar as diferentes situações em que o experimento é realizado.

Uma questão importante neste tipo de análise é a definição do número de componentes k da mistura.

Neste Capítulo descrevemos o modelo com mistura finita, $k < \infty$, e em seguida descrevemos uma modelagem equivalente ao modelo de misturas de processo Dirichlet, considerando um modelo com mistura de distribuições com o número de componentes k tendendo para o infinito, $k \to \infty$. Aplicamos esta modelagem na análise da expressão gênica com o objetivo de identificar grupos de genes com níveis de expressão similares.

5.1 Introdução

Seja $\mathbf{y} = (y_1, y_2, ..., y_n)$ uma amostra observada de tamanho n de um vetor aleatório $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ proveniente de uma população com k sub-populações (componentes), com função densidade de probabilidade, no caso da variável assumir valores contínuos, ou função de probabilidade no caso da variável assumir valores discretos, $f(y_i|\theta_j)$, para $i = 1, 2, ..., n$ e $j = 1, 2, ..., k$ com $k < \infty$. Isto é, $\mathbf{y} = (y_1, y_2, ..., y_n)$ é uma amostra obtida

de uma mistura finita de distribuições, com função densidade ou de probabilidade dada por

$$p\left(\mathbf{y}|\theta\right) = \sum_{j=1}^{k} w_j f\left(y_i|\theta_j\right) \tag{5.1}$$

onde $\theta = (\theta_1, \theta_2, ..., \theta_k)$ é o vetor de parâmetros, $f\left(y_i|\theta_j\right)$ é a densidade da j-ésima componente com parâmetro(s) θ_j e $w = (w_1, w_2, ..., w_k)$ são as proporções associadas as componentes da mistura, com $0 < w_j < 1$, sujeito a restrição $\sum_{j=1}^{k} w_j = 1$, para $i = 1, 2, ..., n$, $j = 1, 2, ..., k$ e $k < \infty$.

Logo, temos que a função de verossimilhança é dada por

$$L\left(\theta|\mathbf{y}\right) = \prod_{i=1}^{n} p\left(y_i|\theta\right) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f\left(y_i|\theta_j\right). \tag{5.2}$$

Note que, pela função de verossimilhança, dada uma observação y_i não identificamos previamente de qual componente esta observação é proveniente. Ou seja, uma questão sobre a utilização de um modelo com mistura, e a respeito da classificação, isto é, da determinação da componente geradora da observação y_i, para $i = 1, 2, ..., n$.

Esta classificação é explicada por Diebolt e Robert (1994), como sendo uma estrutura "oculta", onde cada observação é associada a um indicador não observado que indica de qual componente a observação foi gerada e afirmam que o modelo com mistura como em (5.1) pode ser escrito em termos desta estrutura "oculta". Isto é possível, se introduzimos um vetor de variáveis latentes $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$ tal que, cada variável latente $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ik})$, $i = 1, 2, ..., n$, tem dimensionalidade k e indica a componente geradora da observação y_i da seguinte forma

$$Z_{ij} = \begin{cases} 1, \text{ se a observação } y_i \text{ é gerada pela componente } j \\ 0, \text{ caso contrário} \end{cases}, \tag{5.3}$$

com Z_i sujeita a restição $\sum_{j=1}^{k} Z_{ij} = 1$, para $i = 1, 2, ..., n$.

Dessa forma, considerando $\mathbf{Y}$ e $\mathbf{Z}$ independentes, a função de verossimilhança (5.2)

pode ser escrita como

$$L\left(\theta|\mathbf{y},\mathbf{z}\right)=\prod_{i=1}^{n}\prod_{j=1}^{k}\left[w_{j}f\left(y_{i}|\theta_{j}\right)\right]^{z_{ij}},$$

ou seja, a introdução do conjunto de variáveis latentes $\mathbf{Z}$ simplifica a função de verossimilhança.

Utilizando a abordagem bayesiana, dada a distribuição *a priori* para o(s) parâmetro(s) θ_j das componente, temos uma simplificação da distribuição *a posteriori* e do desenvolvimento das análises.

O interesse agora é verificar de qual distribuição de probabilidades a variável aleatória Y_i é proveniente, através de Z_i, $i=1,2,...,n$. Uma maneira de identificar a origem de Y_i, através de Z_i, e de estimar os coeficiente $w_1,w_2,...w_k$ e os parâmetros $\theta_1,\theta_2,...,\theta_k$ das componentes, utilizando as observações $\mathbf{y}=(y_1,y_2,...,y_n)$, é através da utilização de métodos iterativos tais como o método *Gibbs sampling* (ver Gelfand e Smith, 1990), o método *Metropolis Hastings* (ver Hastings, 1970) e o algoritmo EM (ver, por exemplo, Celeux e Diebolt, 1985).

5.2 Modelo equivalente ao modelo MPD

Utilizando o modelo com mistura em (5.1), podemos obter um modelo equivalente ao modelo de misturas de processos Dirichlet, considerando:

- $\phi=(\phi_1,\phi_2,...,\phi_k)$ um conjunto de "grupos latentes" composto por observações similares, de forma que as observações pertencentes ao "grupo latente" ϕ_j é modelada por uma densidade comum $f(\theta_j)$, para $j=1,2,...,k$. Isto é, cada "grupo latente" é composto por observações geradas de uma mesma distribuição de probabilidades;

- $\mathbf{Z}=(Z_1,Z_2,...,Z_n)$ um conjunto de variáveis latentes, tal que cada $Z_i=(Z_{i1},Z_{i2},...Z_{ik})$ tem dimensionalidade k e indica o "grupo latente" que a observação y_i pertence, da seguinte forma

$$Z_{ij}=\begin{cases}1, \text{ se a observação } y_i \text{ pertence ao "grupo latente" } \phi_j\\ 0, \text{ caso contrário}\end{cases},$$

com n_j representando a quantidade de observações pertencentes ao "grupo latente" ϕ_j,

para $i = 1, 2, ..., n$ e $j = 1, 2, ..., k$;

- e considerando a seguinte modelagem hierárquica

$$\begin{aligned} y_i|c, \phi, \theta_j &\sim f(\theta_j) \\ Z_1, Z_2, ..., Z_n|w_1, w_2, ..., w_k &\sim Multinomial(1; w_1, w_2, ..., w_k) \\ \theta_j &\sim G_0 \\ w_1, w_2, ..., w_k &\sim Dirichlet\left(\frac{M}{k}, \frac{M}{k}, ..., \frac{M}{k}\right) \end{aligned} \tag{5.4}$$

com o número de componentes k da mistura tendendo para o infinito, $k \to \infty$, onde G_0 é uma distribuição paramétrica especificada, conhecida, com densidade g_0 e M é uma constante conhecida (como definida na seção 4.3), para $j = 1, 2, ..., k$ e $k \to \infty$.

Portanto de (5.4) temos que dadas as proporções associadas as componentes da mistura, a distribuição *a priori* para as variáveis latentes é dada por uma distribuição multinomial. Para as proporções $w_1, w_2, ..., w_k$ associadas às componentes é considerada uma distribuição *a priori* usual Dirichlet, com parâmetros $\frac{M}{k}$, que é aproximadamente zero quando $k \to \infty$. Para o(s) parâmetro(s) θ_j de cada componente associada as observações de um "grupo latente", é considerada uma distribuição *a priori* G_0, conhecida, para $j = 1, 2, ..., k$ e $k \to \infty$.

MacEachern e Müller (1998) chamam cada "grupo latente" ϕ_j de *cluster*, devido ϕ_j formar um "aglomerado" ou um "*cluster*" de observações similares.

Dessa forma, temos que com a modelagem (5.4) cada "grupo latente" ϕ_j determina um *cluster* de observações similares. Com cada *cluster* sendo modelado por uma densidade comum f com parâmetro(s) θ_j. Com isso, *clusters* de observações similares e observações que não são similares com qualquer outra observação são facilmente detectadas.

De modo a facilitar a demonstração da equivalência entre (5.4) com o modelo de misturas de processo Dirichlet em (4.16) e viabilizar o desenvolvimento de um procedimento de simulação, considere $k < \infty$ e $\mathbf{c} = (c_1, c_2, ..., c_n)$ um conjunto de variáveis indicadoras, tal que

$$c_i = \begin{cases} j, \text{ se a variável latente } Z_{ij} = 1 \\ 0, \text{ caso contrário} \end{cases},$$

isto é, $c_i = j$ indica que a observação y_i pertence ao *cluster* ϕ_j, com n_j representando

a quantidade de observações pertencentes ao *cluster* ϕ_j (ou a quantidade de c_i's $= j$), para $j = 1, 2, ..., k$. Sendo necessário escrever a probabilidade *a priori* para c_i dado $c_1, c_2, ..., c_{i-1}$, para $i = 1, 2, ..., n$.

Dessa forma, temos que

$$\begin{aligned} P(w_1, w_2, ..., w_k | c_1, c_2, ..., c_n) &\propto \prod_{j=1}^{k} w_j^{n_j} \prod_{j=1}^{k} w_j^{\frac{M}{k}-1} \propto \prod_{j=1}^{k} w_j^{n_j+\frac{M}{k}-1} \\ &= Dirichlet\left(n_1 + \frac{M}{k}, n_2 + \frac{M}{k}, ..., n_k + \frac{M}{k}\right). \end{aligned} \tag{5.5}$$

A probabilidade de uma nova observação y_{n+1} pertencer ao *cluster* ϕ_j dado o *cluster* das n primeiras observações, isto é, a probabilidade da variável indicadora c_{n+1} ser igual a j, $c_{n+1} = j$, dado o valor das n primeiras variáveis indicadoras, é dado por

$$\begin{aligned} P(c_{n+1} = j | c_1, ..., c_n) &= \int_w P(c_{n+1} = j | w_1, ..., w_k) P(w_1, ..., w_k | c_1, ..., c_n) \, dw \\ &= \int_w w_j P(w_1, w_2, ..., w_k | c_1, c_2, ..., c_n) \, dw \\ &= \int_w w_j Dirichlet\left(n_1 + \frac{M}{k}, n_2 + \frac{M}{k}, ..., n_k + \frac{M}{k}\right) dw \\ &= E(w_j) \\ &= \frac{n_j + \frac{M}{k}}{M + n} \end{aligned} \tag{5.6}$$

para $j = 1, 2, ..., k$ e $k < \infty$.

Note que (5.6) é obtida marginalizando sobre $w = (w_1, w_2, ..., w_k)$.

Por analogia com (5.6), temos que a probabilidade condicional de c_i dado $c_1, c_2, ..., c_{i-1}$ é dada por

$$P(c_i = j | c_1, c_2, ..., c_{i-1}) = \frac{n_j + \frac{M}{k}}{M + i - 1} \tag{5.7}$$

para $i = 1, 2, ..., n$ e $j = 1, 2, ..., i - 1$.

Considerando agora (5.7) com o número de componente k tendendo para o infinito, $k \to \infty$, temos que

$$P(c_i = j | c_1, c_2, ..., c_{i-1}) = \frac{n_j}{M + i - 1} \tag{5.8}$$

para $i = 1, 2, ..., n$ e $j = 1, 2, ..., i - 1$.

A probabilidade de $c_i \neq j$, é dada por

$$\begin{aligned} P\left(c_i \neq j | c_1, c_2, ..., c_{i-1}\right) &= 1 - \sum_{j=1}^{i-1} \frac{n_j}{M+i-1} \\ &= 1 - \frac{1}{M+i-1} \sum_{j=1}^{i-1} n_j \\ &= \frac{M+i-1-\sum_{j=1}^{i-1} n_j}{M+i-1} \\ &= \frac{M}{M+i-1} \end{aligned} \tag{5.9}$$

ou seja, (5.9) é a probabilidade *a priori* de y_i não pertencer a nenhum dos *clusters* das $i-1$ primeiras observações. Sendo necessário a "criação"de um novo *cluster* para y_i associando a este uma densidade f com parâmetros(s) gerados da distribuição G_0, para $i = 1, 2, ..., n$ e $j = 1, 2, ...i - 1$.

Em resumo, temos que a probabilidade *a priori* para c_i dado $c_1, c_2, ..., c_{i-1}$, com $k \to \infty$, é dada por

$$\begin{aligned} P\left(c_i = j | c_1, c_2, ...c_{i-1}\right) &= \frac{n_j}{M+i-1} \\ P\left(c_i \neq j | c_1, c_2, ..., c_{i-1}\right) &= \frac{M}{M+i-1} G_0 \end{aligned} \tag{5.10}$$

para $i = 1, 2, ..., n$ e $j = 1, 2, ..., i - 1$.

Assim, temos que (5.10) representa a probabilidade *a priori* para $\theta_1, \theta_2, ..., \theta_k$, através de $\mathbf{c} = (c_1, c_2, ..., c_n)$, como sendo uma mistura de distribuições, com uma parte continua, gerada segundo uma distribuição paramétrica G_0 e uma parte discreta, que repete os valores das componentes anteriores. Com isso, temos que alguns valores de $\theta_1, \theta_2, ..., \theta_k$ podem ser iguais e cada observação associada a um mesmo θ_j pertence a um mesmo *cluster* é modelada por uma densidade comum $f(\theta_j)$, para $j = 1, 2, ..., i - 1$.

Portanto, temos que o limite do modelo (5.4) para $k \to \infty$, é equivalente ao modelo de misturas de processo Dirichlet em (4.16), devido a correspondência entre as probabilidades condicionais para θ_i em (4.17) e (5.10) (ver, MacEachern e Müller, 1998 e Neal 1998).

Como em (4.18), temos que (5.10) condicionado sobre todas variáveis indicadoras in-

dexadas de 1 a n exceto a i-ésima (denotamos este vetor por $c_{-1} = (c_1, c_2, ..., c_{i-1}, c_{i+1}, ..., c_n)$), é dada por

$$\begin{aligned} P\left(c_i = j | c_{-i}\right) &= \frac{n_{-i,j}}{M+n-1} \\ P\left(c_i \neq j | c_{-i}\right) &= \frac{M}{M+n-1} G_0 \end{aligned} \tag{5.11}$$

onde $n_{-i,j}$ é a quantidade de $c_{i'}$, para $i' \neq i$, que são iguais a j $(c_{i'} = j)$, para $i = 1, 2, ..., n$ e $j = 1, 2, ..., k$ e $k \to \infty$.

Como feito em (4.21), combinando (5.11) com a função de verossimilhança $L(\theta_j | y_i) = f\left(y_i | \theta_j\right)$ (baseada apenas na observação y_i, isto é, a função de verossimilhança e o valor da densidade f com parâmetro(s) θ_j no ponto y_i) temos que a probabilidade condicional de c_i dado c_{-i}, é dada por

$$\begin{aligned} P\left(c_i = j | c_{-i}\right) &= b \frac{n_{-i,j}}{M+n-1} f\left(y_i | \theta_j\right) \\ P\left(c_i \neq j | c_{-i}\right) &= b \frac{M}{M+n-1} q_{0i} P\left(\theta_j | y_i\right) \end{aligned} \tag{5.12}$$

onde,

$$q_{0i} = \int g_0(\theta_j) f\left(\theta_j\right) d\theta_j, \tag{5.13}$$

e

$$P\left(\theta_j | y_i\right) = \frac{g_0(\theta_j) f\left(y_i | \theta_j\right)}{\int g_0(\theta_j) f\left(y_i | \theta_j\right) d\theta_j} \propto g_0(\theta_j) f\left(y_i | \theta_j\right) \tag{5.14}$$

é a distribuição *a posteriori* de θ_j dado que $c_i \neq j$ e y_i, e

$$b = \left(\frac{M q_{0i} + n_{-i,j} f\left(y_i | \theta_j\right)}{M+n-1} \right)^{-1} \tag{5.15}$$

é a constante normalizadora apropriada, para $i = 1, 2, ..., n$, $j = 1, 2, ..., k$ e $k \to \infty$.

Definidas as probabilidades condicionais de $c_i | c_{-i}, y_i$, temos que dado os valores de $\mathbf{c} = (c_1, c_2, ..., c_n)$, temos formado uma estrutura de *clusters* $\phi_1, \phi_2, ...$ de observações similares. Para cada *cluster* ϕ_j temos uma função de verossimilhança associada, que é

dada por

$$L\left(\theta_j|\mathbf{y}_{\phi_j}\right) = \prod_{i=1}^{n_j} f\left(\mathbf{y}_{\phi_j}|\theta_j\right) \tag{5.16}$$

onde $\mathbf{y}_{\phi_j} = \begin{pmatrix} y_1, y_2, ..., y_{n_j} \\ \phi_j \quad \phi_j \qquad \phi_j \end{pmatrix}$ representa as n_j observações pertencentes ao *cluster* ϕ_j, $j = 1, 2, ..., k$ e $k \to \infty$.

Dessa forma, temos que a distribuição *a posteriori* do(s) parâmetro(s) θ_j da distribuição de probabilidades associada ao *cluster* ϕ_j dado as observações $\mathbf{y}_{\phi_j}$, é dada por

$$P\left(\theta_j|\mathbf{y}_{\phi_j}\right) \propto L\left(\theta_j|\mathbf{y}_{\phi_j}\right) g_0(\theta_j). \tag{5.17}$$

Definidas as probabilidades condicionais de $c_i|c_{-i}, y_i$ e a distribuição *a posteriori* dos parâmetros associados a cada *cluster* formado, dado os valores de $c_1, c_2, ..., c_n$, é proveitoso deduzir um método de amostragem para o modelo com mistura infinita equivalente ao modelo MPD.

5.2.1 *Gibbs Sampling* para modelos conjugados

Quando k tende para o infinito, $k \to \infty$, não podemos representar explicitamente o número infinito de *clusters* ϕ_j, $j = 1, 2, ...$ Dessa forma, para viabilizar o método representamos somente os ϕ_j que são associados a pelo menos uma observação (ver Neal, 1998). Isto é, fazemos $j = 1, 2, ..., k^*$, onde k^* e a quantidade de *clusters* associado a pelo menos uma observação. Dessa forma a amostragem se torna equivalente a amostragem de um modelo com mistura finita de distribuições.

Como mencionado na seção 4.4, se as densidades f e g_0 são conjugadas a integração envolvida no cálculo de q_{0i} pode ser feita analiticamente e a amostragem de θ_i se torna computacionalmente fácil. Logo, podemos utilizar o método *Gibbs sampling* para se obter amostras das distribuições condicionais de θ_i, para $i = 1, 2, ..., n$, como descrito a seguir:

Algoritmo 2: Considere um estado da cadeia de *Markov* consistindo de $c_1, c_2, ..., c_n$ e $\phi = \left(\phi_j : j \in \{1, 2, ..., k^*\}\right)$. Repetidamente amostre como a seguir, até obter convergência para o número de componentes k^* e para os parâmetros das distribuições associadas aos *clusters*:

- Tire um valor para $c_i|c_{-i}, y_i$ com probabilidades definidas pela equação (5.12);

i) se o valor de c_i é associado a algum *cluster* já presente no estado da cadeia de *Markov*, então para todo $\mathbf{c} = (c_1, c_2, ..., c_n)$ e $j \in \{1, 2, ..., k^*\}$ tal que $c_i = j$, gere um valor para $\theta_j|\mathbf{y}_{\phi_j}$ de sua distribuição *a posteriori* como em (5.17), para $i = 1, 2, ..., n$;

ii) se o valor de c_i não é associado a nenhum dos j *clusters* das outras observações, para $j = 1, 2, ..., k^*$, faça $k^* = k^* + 1$ e crie um novo *cluster* para y_i associando a este uma distribuição de probabilidades com parâmetro(s) θ_{k^*} gerados da distribuição *a posteriori* $P(\theta_{k^*}|y_i)$, como em (5.14). Adicione esta nova componente ao estado da cadeia de *Markov*.

iii) se algum *cluster* ϕ_j, $j = 1, 2, ..k^*$, presente no estado da cadeia de *Markov* não possuir nenhuma observação, então faça $k^* = k^* - 1$ e retire este *cluster* do estado da cadeia de *Markov*.

Este método é utilizado por Bush e MacEachern (1996) e por West, Müller e Escobar (1994). Como no caso do primeiro método *Gibbs sampling* descrito (algoritmo 1 - seção 4 do Capítulo 4), esta abordagem se torna viável somente se podemos calcular q_{0i} e amostrar de $P(\theta_j|y_i)$, para $i = 1, 2, ..., n$, $j = 1, 2, ..., k^*$. Isto geralmente, é possível quando as densidades f e g_0 são conjugadas. Para procedimentos de simulação, onde f e g_0 não são conjugadas ver MacEachern e Müller, 1998 e Neal, 1998.

Como temos n observações, o número de *clusters* formados pode ser no maximo n, $k^* = n$, ou seja $\phi = (\phi_1, \phi_2, ..., \phi_n)$. Dessa forma, sugerimos para facilitar a implementação do algoritmo 2, considerar o número de componentes $k^* = n$ no método *Gibbs sampling*, em vez de considerar k^* variando (algoritmo 2), e seguir os passos do algoritmo 2.1 abaixo.

Algoritmo 2.1: Considere um estado da cadeia de *Markov* consistindo de $c_1, c_2, ..., c_n$ e $\phi = \left(\phi_j : j \in \{1, 2, ..., n\}\right)$. Repetidamente amostre como a seguir, até obter convergência:

∘ Tire um valor para $c_i|c_{-i}, y_i$ com probabilidades definidas pela equação (5.12), ou seja

$$\begin{array}{lll} \text{Amostre } c_1 \text{ de } & \left[c_1|c_2 = c_2^{p-1}, c_3 = c_3^{p-1}, ..., c_n = c_n^{p-1}\right] \\ \text{Amostre } c_2 \text{ de } & \left[c_2|c_1 = c_1^{p}, c_3 = c_3^{p-1}, ..., c_n = c_n^{p-1}\right] \\ \vdots & \vdots & \vdots \\ \text{Amostre } c_n \text{ de } & \left(c_n|c_1 = c_1^{p}, c_2 = c_2^{p}, ..., c_{n-1} = c_{n-1}^{p}\right) \end{array}$$

- Dada a p-ésima seqüência gerada pelo método *Gibbs sampling*, temos um conjunto de valores para as variáveis indicadoras $\mathbf{c}^p = (c_1^p, c_2^p, ..., c_n^p)$, que podem ser $c_i^p = j$ para $j = \{1, 2, ..., n\}$ ou $c_i^p = n+1$ se a observação y_i não é associada a nenhum dos n *clusters*. Assim:

i) se c_i^p é associado a algum *cluster* com pelo menos uma observação, isto é, $c_i^p = j$ para algum $j \in \{1, 2, ..., n\}$ tal que $n_{-i,j} > 0$, então y_i pertence ao *cluster* ϕ_j, que agora tem mais uma observação, $n_j = n_j + 1$. Atualize θ_j da distribuição *a posteriori* como em (5.17);

ii) se c_i^p não é associado a nenhum dos n *clusters*, isto é, $c_i^p = n+1$, aloque a observação y_i para um *cluster* que não possui nenhuma observação, ou seja, faça $c_i^p = j$ para algum $j \in \{1, 2, ..., n\}$ tal que $n_{-i,j} = 0$. Feito isso, associamos a y_i um *cluster* ϕ_j que não possuía nenhuma observação e agora possui uma, $n_j = 1$, e associamos a este *cluster* uma nova distribuição de probabilidades com parâmetros θ_j gerados da distribuição *a posteriori* (5.14);

iii) para os *clusters* sem nenhuma observação, isto é, $n_j = 0$, gere para o(s) parâmetro(s) θ_j, da distribuição de probabilidades associada a este(s) *cluter(s)*, valores da distribuição *a priori* G_0.

Se P for todas as seqüência gerada pelo procedimento *Gibbs sampling*, então a observação y_i será associada ao *cluster* ϕ_j se a proporção de vezes que y_i foi associado ao *cluster* ϕ_j for maior que a proporção de vezes que a observação y_i foi associado a qualquer outro *cluster* ao final das P seqüências geradas pelo método *Gibbs sampling*. Dessa forma, a convergência para o número de *cluster* se da pela quantidade de *clusters*, dentre os n, que possuem pelo menos uma observação o final das P seqüências geradas pelo método *Gibbs sampling*.

Caso 3: Modelo com mistura infinita com variâncias desconhecidas

Para ilustrar a aplicação da modelagem com mistura infinita que é equivalente ao modelo MPD e do algoritmo 2.1, consideramos o mesmo procedimento descrito para o caso 2: Variâncias desconhecidas, feita na subseção 4.4. Isto é, consideramos que as observações $y_1, y_2, ..., y_n$ sejam distribuídas como uma mistura de distribuições normais com médias μ_i e variâncias σ_i^2 (desconhecidas), isto é $y_i \sim N(\mu_i, \sigma_i^2)$, e a seguinte modelagem equivalente

ao modelo MPD, como em (5.4)

$$\begin{aligned} y_i|c,\phi,\mu_j,\sigma_j^2 &\sim N\left(\mu_j,\sigma_j^2\right) \\ Z_1,Z_2,...,Z_n|w_1,w_2,...,w_k &\sim Multinomial\left(1;w_1,w_2,...,w_n\right) \\ \mu_j,\sigma_j^2|G_0 &\sim G_0 \\ w_1,w_2,...,w_n &\sim Dirichlet\left(\frac{M}{k},\frac{M}{k},...,\frac{M}{k}\right) \end{aligned} \tag{5.18}$$

com

$$M = a \text{ (constante)}, G_0\left(\mu_j|\sigma_j^2\right) = N\left(0,\frac{\sigma_j^2}{\lambda}\right), G_0\left(\sigma_j^2\right) = IG\left(\frac{\alpha}{2},\frac{\beta}{2}\right),$$

ou seja,

$$G_0\left(\mu_j,\sigma_j^2\right) = N\left(0,\frac{\sigma_j^2}{\lambda}\right) IG\left(\frac{\alpha}{2},\frac{\beta}{2}\right) \text{ e } g_0\left(\mu_j,\sigma_j^2\right) = f_{N\left(0,\frac{\sigma_j^2}{\lambda}\right)}\left(\mu_j\right) f_{IG\left(\frac{\alpha}{2},\frac{\beta}{2}\right)}\left(\sigma_j^2\right)$$

para a, λ, α e β conhecidos e fixos, e

$$c_i = \begin{cases} j, \text{ se a variável latente } Z_{ij} = 1 \\ 0, \text{ caso contrário} \end{cases},$$

para $i = 1,2,..,n$ e $j = 1,2,...,k^*$.

A probabilidade condicional para $c_i|c_{-i},y_i$, como em (5.12), é dada por

$$\begin{aligned} P\left(c_i = j|c_{-i}\right) &= \frac{n_{-i,j}}{M+n-1} f_{N\left(\mu_i,\sigma_i^2\right)}\left(y_i\right) \\ P\left(c_i \neq j|c_{-i}\right) &= \frac{M}{M+n-1} q_{0i} P\left(\mu_j,\sigma_j^2|y_i\right) \end{aligned} \tag{5.19}$$

onde $f_{N\left(\mu_i,\sigma_i^2\right)}\left(y_i\right)$ é o valor da densidade da distribuição normal com média μ_i e variância σ_i^2 no ponto y_i, e

$$q_{oi} = f_{t_\alpha\left(0,\frac{B(\lambda+1)}{\alpha\lambda}\right)}\left(y_i\right) \tag{5.20}$$

é o valor da densidade da distribuição *t-Student* com média zero, variância $\frac{B(\lambda+1)}{\alpha\lambda}$ e α graus de liberdade, no ponto y_i, e

$$P\left(\mu_j,\sigma_j^2|,y_i\right) = N\left(\frac{y_i}{\lambda+1},\frac{\sigma_j^2}{\lambda+1}\right) IG\left(\frac{\alpha+2}{2},\frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right) \tag{5.21}$$

com distribuições condicionais dadas por

$$P\left(\mu_j|, \sigma_j^2, y_i\right) = N\left(\frac{y_i}{\lambda+1}, \frac{\sigma_j^2}{\lambda+1}\right) \tag{5.22}$$

e

$$P\left(\sigma_j^2|y_i\right) = IG\left(\frac{\alpha+2}{2}, \frac{\beta\left(\lambda+1\right)+\lambda y_i^2}{2\left(\lambda+1\right)}\right), \tag{5.23}$$

onde $N\left(\frac{y_i}{\lambda+1}, \frac{\sigma_j^2}{\lambda+1}\right)$ e $IG\left(\frac{\alpha+2}{2}, \frac{\beta(\lambda+1)+\lambda y_i^2}{2(\lambda+1)}\right)$ representam, respectivamente, a distribuição normal com média $\frac{y_i}{\lambda+1}$ e variância $\frac{\sigma_j^2}{\lambda+1}$ e a distribuição gama inversa com parâmetro de forma $\frac{\alpha+2}{2}$ e parâmetro de escala $\frac{\beta(\lambda+1)+\lambda y_i^2}{2(\lambda+1)}$ (ver os cálculos de q_{oi} e $P\left(\mu_i|\sigma_i^2, y_i\right)$ em caso 2: variâncias desconhecidas, na subseção 4.4.1 da seção 4.4), para $i = 1, 2, ..., n$ e $j = 1, 2, ..., k^*$.

A constante normalizadora apropriada, como em (5.15), é dada por

$$b = \left(\frac{n_{-i,j} f_{N\left(\mu_i, \sigma_i^2\right)}\left(y_i\right) + M f_{t_\alpha\left(0, \frac{B(\lambda+1)}{\alpha\lambda}\right)}\left(y_i\right)}{M+n-1}\right)^{-1},$$

para $i = 1, 2, ..., n$ e $j = 1, 2, ..., k^*$.

Como sugerido no algoritmo 2.1, considerando $k^* = n$, dados os valores de $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ definidas como em (4.19), temos as quantidades $(n_1, n_2, ..., n_n)$ pertencentes aos *clusters* $\phi_1, \phi_2, ..., \phi_n$, respectivamente. Para cada *cluster* ϕ_j, como em (4.16), temos uma função de verossimilhança associada, que é dada por

$$L\left(\theta_j|\mathbf{y}_{\phi_j}\right) = \prod_{i=1}^{n_J} f\left(\mathbf{y}_{\phi_j}|\theta_j\right) \propto \left(\sigma_j^2\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}\left(\underset{\phi_j}{y_i} - \mu_j\right)^2\right\}$$

que pode ser escrita sob a forma,

$$L\left(\theta_j|\mathbf{y}_{\phi_j}\right) \propto \left(\sigma_j^2\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\sigma_j^2}\left[\left(\underset{\phi_j}{\bar{y}} - \mu_j\right)^2 + \left(n_j-1\right)\underset{\phi_j}{s^2}\right]\right\} \tag{5.24}$$

onde $\mathbf{y}_{\phi_j} = \left(\underset{\phi_j}{y_1}, \underset{\phi_j}{y_2}, ..., \underset{\phi_j}{y_{n_j}}\right)$ representa as n_j observações pertencentes ao *cluster* ϕ_j e $\underset{\phi_j}{\bar{y}}$ e $\underset{\phi_j}{s^2}$ é a média e variância amostral das n_j observações do *cluster* ϕ_j, respectivamente, para

$j = 1, 2, ..., n$.

Logo, como em (4.17) a distribuição *a posteriori* dos parâmetros μ_j, σ_j^2, da distribuição normal associada ao *cluster* ϕ_j dado $\mathbf{c} = (c_1, c_2, ..., c_n)$ e as observações $\mathbf{y}_{\phi_j}$, é dada por

$$
\begin{aligned}
P\left(\mu_j, \sigma_j^2 | \mathbf{y}_{\phi_j}\right) &\propto L\left(\mu_j, \sigma_j^2 | \mathbf{c}, \phi_j, \mathbf{y}_{\phi_j}\right) g_0(\mu_j, \sigma_j^2) \\
&\propto L\left(\mu_j, \sigma_j^2 | \mathbf{c}, \phi_j, \mathbf{y}_{\phi_j}\right) f_{N\left(0, \frac{\sigma_j^2}{\lambda}\right)}\left(\mu_j\right) f_{IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)}\left(\sigma_j^2\right) \\
&\propto \left(\sigma_j^2\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\sigma_j^2}\left[n_j\left(\bar{y}_{\phi_j} - \mu_j\right)^2 + (n_j - 1) s_{\phi_j}^2\right]\right\} \\
&\quad \left(\sigma_j^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda}{2\sigma_j^2}\mu_j^2\right\} \left(\sigma_j^2\right)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_j^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_j^2}\left[n_j\left(\bar{y}_{\phi_j} - \mu_j\right)^2 + (n_j - 1) s_{\phi_j}^2\right] - \frac{\lambda}{2\sigma_j^2}\mu_j^2\right\} \\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_j^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_j^2}\left[n_j\left(\bar{y}_{\phi_j} - \mu_j\right)^2 + \lambda\mu_j^2\right]\right\} \\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_j^2} - \frac{1}{2\sigma_j^2}(n_j - 1) s_{\phi_j}^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_j^2}\left[n_j\left(\bar{y}_{\phi_j}^2 - 2\mu_j \bar{y}_{\phi_j} + \mu_j^2\right) + \lambda\mu_j^2\right]\right\} \\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_j^2} - \frac{(n_j - 1) s_{\phi_j}^2}{2\sigma_j^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_j^2}\left[n_j \bar{y}_{\phi_j}^2 - 2\mu_j n_j \bar{y}_{\phi_j} + n_j\mu_j^2 + \lambda\mu_j^2\right]\right\} \\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_j^2} - \frac{(n_j - 1) s_{\phi_j}^2}{2\sigma_j^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_j^2}\left[(n_j + \lambda)\mu_j^2 - 2\mu_j n_j \bar{y}_{\phi_j}\right]\right\} \\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma_j^2} - \frac{(n_j - 1) s_{\phi_j}^2}{2\sigma_j^2} - \frac{n_j \bar{y}_{\phi_j}^2}{2\sigma_j^2}\right\}
\end{aligned}
\tag{5.25}
$$

$$
\begin{aligned}
&\propto \exp\left\{-\frac{n_j+\lambda}{2\sigma_j^2}\left[\mu_j^2-2\mu_j\left(\frac{n_j\,\bar{y}_{\phi_j}}{n_j+\lambda}\right)\right]\right\}\\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_j^2}-\frac{(n_j-1)\,s^2_{\phi_j}}{2\sigma_j^2}-\frac{n_j\bar{y}^2_{\phi_j}}{2\sigma_j^2}\right\}\\
&\propto \exp\left\{-\frac{1}{2\frac{\sigma_j^2}{n_j+\lambda}}\left[\mu_j-\left(\frac{n_j\,\bar{y}_{\phi_j}}{n_j+\lambda}\right)\right]^2\right\}\\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)}\exp\left\{-\frac{\beta}{2\sigma_j^2}-\frac{(n_j-1)\,s^2_{\phi_j}}{2\sigma_j^2}-\frac{n_j\bar{y}^2_{\phi_j}}{2\sigma_j^2}+\frac{\left(n_j\,\bar{y}_{\phi_j}\right)^2}{2\sigma_j^2\,(n_j+\lambda)}\right\}\\
&\propto \exp\left\{-\frac{1}{2\frac{\sigma_j^2}{n_j+\lambda}}\left[\mu_j-\left(\frac{n_j\,\bar{y}_{\phi_j}}{n_j+\lambda}\right)\right]^2\right\}\\
&\quad \left(\sigma_j^2\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)}\exp\left\{-\frac{1}{\sigma_j^2}\left[\frac{\beta\,(n_j+\lambda)+(n_j+\lambda)\,(n_j-1)\,s^2_{\phi_j}+n_j\lambda\bar{y}^2_{\phi_j}}{2\,(n_j+\lambda)}\right]\right\}\\
&= N\left(\frac{n_j\,\bar{y}_{\phi_j}}{n_j+\lambda},\frac{\sigma_j^2}{n_j+\lambda}\right)IG\left(\frac{\alpha+n_j+1}{2},\frac{\beta\,(n_j+\lambda)+(n_j+\lambda)\,(n_j-1)\,s^2_{\phi_j}+n_j\lambda\bar{y}^2_{\phi_j}}{2\,(n_j+\lambda)}\right)
\end{aligned}
$$

onde $N\left(\frac{n_j\,\bar{y}_{\phi_j}}{n_j+\lambda},\frac{\sigma_j^2}{n_j+\lambda}\right)$ e $IG\left(\frac{\alpha+n_j+1}{2},\frac{\beta(n_j+\lambda)+(n_j+\lambda)(n_j-1)s^2_{\phi_j}+n_j\lambda\,\bar{y}^2_{\phi_j}}{2(n_j+\lambda)}\right)$ representam, respectivamente, a distribuição normal com média $\frac{n_j\,\bar{y}_{\phi_j}}{n_j+\lambda}$ e variância $\frac{\sigma_j^2}{n_j+\lambda}$ e a distribuição gama inversa com parâmetro de forma $\frac{\alpha+n_j+1}{2}$ e parâmetro de escala $\frac{\beta(n_j+\lambda)+(n_j+\lambda)(n_j-1)s^2_{\phi_j}+n_j\lambda\,\bar{y}^2_{\phi_j}}{2(n_j+\lambda)}$, para $i=1,2,...,n$ e $j=1,2,...,n$.

As distribuições condicionais utilizadas no método *Gibbs Samplig* para estimação dos parâmetros, são dadas por

$$
P\left(\mu_j|\sigma_j^2,\mathbf{y}_{\phi_j}\right)=N\left(\frac{n_j\,\bar{y}_{\phi_j}}{n_j+\lambda},\frac{\sigma_j^2}{n_j+\lambda}\right) \tag{5.26}
$$

e

$$P\left(\sigma_j^2|\mathbf{y}_{\phi_j}\right) = IG\left(\frac{\alpha+n_j+1}{2}, \frac{\beta\left(n_j+\lambda\right)+\left(n_j+\lambda\right)\left(n_j-1\right)\underset{\phi_j}{s^2}+n_j\lambda\underset{\phi_j}{\bar{y}^2}}{2\left(n_j+\lambda\right)}\right) \tag{5.27}$$

para $j = 1, 2, ..., n$.

Definido as probabilidades condicionais de $c_i|c_{-i}, y_i$, e a distribuição *a posteriori* de $\mu_j, \sigma_j^2|\mathbf{y}_{\phi_j}$, associado ao *cluster* ϕ_j, inicializamos o método *Gibbs sampling* fazendo $c_i^0 = i$, ou seja, inicialmente cada observação em um *cluster*, $\mu_i = y_i$ e $\sigma_i^2 = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$, para $i = 1, 2, ..., n$, e a *p-ésima* sequência gerada pelo método *Gibbs sampling* é feita da seguinte maneira,

$$\begin{array}{lll} \text{Amostre } c_1 \text{ de } \left[c_1|c_2 = c_2^{p-1}, c_3 = c_3^{p-1}, ..., c_n = c_n^{p-1}\right] & & \\ \text{Amostre } c_2 \text{ de } \left[c_2|c_1 = c_1^{p}, c_3 = c_3^{p-1}, ..., c_n = c_n^{p-1}\right] & & \\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots & & \vdots \\ \text{Amostre } c_n \text{ de } \left[c_n|c_1 = c_1^{p}, c_2 = c_2^{p}, ..., c_{n-1} = c_{n-1}^{p}\right] & & \end{array} \tag{5.28}$$

com probabilidades definidas como em (5.19);

- Dada a p-ésima sequência gerada pelo método *Gibbs Samplig*, temos um conjunto de valores para as variáveis indicadoras $\mathbf{c}^p = (c_1^p, c_2^p, ..., c_n^p)$, que podem ser $c_i^p = j$ para $j = \{1, 2, ..., n\}$ ou $c_i^p = n+1$ se a observação y_i não é associada a nenhum dos n *clusters*. Assim,

i) se c_i^p é associado a algum *cluster* com pelo menos uma observação, isto é, $c_i^p = j$ para algum $j \in \{1, 2, ..., n\}$ tal que $n_{-i,j} > 0$, então y_i pertence ao *cluster* ϕ_j, que agora tem mais uma observação, $n_j = n_j + 1$. Atualize μ_j e σ_j^2, respectivamente, de (5.26) e (5.27);

ii) se c_i^p não é associado a nenhum dos n *clusters*, isto é, $c_i^p = n+1$, aloque a observação y_i para um *cluster* que não possui nenhuma observação, ou seja, faça $c_i^p = j$ para algum $j \in \{1, 2, ..., n\}$ tal que $n_{-i,j} = 0$. Feito isso, associamos a y_i um *cluster* ϕ_j que não possuía nenhuma observação e agora possui uma, $n_j = 1$, e associamos a este *cluster* uma nova

distribuição normal com parâmetros μ_j e σ_j^2 gerados, respectivamente, das distribuições condicionais (5.22) e (5.23);

iii) Para os *clusters* sem nenhuma observação, isto é, $n_j = 0$, gere para o(s) parâmetro(s) μ_j e σ_j^2, da distribuição normal associada a este(s) *cluster(s)*, valores da distribuição *a priori* G_0, onde

$$G_0\left(\mu_j|\sigma_j^2\right) = N\left(0, \frac{\sigma_j^2}{\lambda}\right) \text{ e } G_0\left(\sigma_j^2\right) = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

Terminada as $p = 1, 2, ..., P$ sequência gerada pelo procedimento *Gibbs sampling*, a observação y_i será associada ao *cluster* ϕ_j se a proporção de vezes que y_i foi associado ao *cluster* ϕ_j for maior que a proporção de vezes que a observação y_i foi associado aos outros *clusters* ao final das P seqüências geradas pelo método *Gibbs sampling*. O número de *cluster* formados será exatamente quantidades de *clusters*, dentre os n, que possuem pelo menos uma observação ao final das P seqüências geradas pelo método *Gibbs sampling*.

Exemplo: Modelo com mistura infinita com variâncias desconhecidas

Para o procedimento de simulação, consideramos os mesmos valores gerados na simulação realizada nos exemplos para os casos 1 e 2, descritros na seção 4.4. Isto é, $y_i \sim N(2, 1)$, para $i = 1, 2, 3, 4, 5$ e aplicamos a modelagem (5.18), com $M = 1$ e os hiperparâmetros foram escolhidos de forma a termos distribuições *a priori* pouco informativa, para isto fizemos $\lambda = 0,01$, $\alpha = 2$ e $\beta = 2$.

Como os cinco valores foram gerados de uma mesma distribuição normal, esperamos que a modelagem com mistura infinita detecte estes cinco valores como pertencentes a um mesmo *cluster*, com parâmetros μ e σ^2 próximos do verdadeiro valor que é 2 e 1, respectivamente.

As probabilidades condicionais de $c_i|c_{-i}, y_i$, como em (5.19), são dadas por

$$\begin{aligned} P\left(c_i = j|c_{-i}\right) &= b\frac{n_{-i,j}}{1+6-1} f_{N\left(\mu_i,\sigma_i^2\right)}\left(y_i\right) = b\frac{n_{-i,j}}{6} f_{N\left(\mu_i,\sigma_i^2\right)}\left(y_i\right) \\ P\left(c_i \neq j|c_{-i}\right) &= b\frac{1}{1+6-1} q_{0i} P\left(\theta_j|y_i\right) = b\frac{1}{6} q_{0i} P\left(\theta_j|y_i\right) \end{aligned}$$

onde

$$q_{oi} = f_{t_\alpha(0,101)}\left(y_i\right)$$

e

$$P\left(\mu_j, \sigma_j^2 | y_i\right) = N\left(\frac{y_i}{1,01}, \frac{\sigma_j^2}{1,01}\right) IG\left(2, \frac{2,02 + (0,01)y_i^2}{2,02}\right)$$

com distribuições condicionais dadas por

$$P\left(\mu_j | \sigma_j^2, y_i\right) = N\left(\frac{y_i}{1,01}, \frac{\sigma_j^2}{1,01}\right)$$

e

$$P\left(\sigma_j^2 | y_i\right) = IG\left(2, \frac{2,02 + (0,01)y_i^2}{2,02}\right).$$

A constante normalizadora apropriada é dada por

$$b = \left(\frac{n_{-i,j} f_{N\left(\mu_i, \sigma_i^2\right)}\left(y_i\right) + f_{t_\alpha(0,101)}\left(y_i\right)}{6}\right)^{-1}.$$

Dados os valores de $\mathbf{c} = (c_1, c_2, ..., c_5)$, temos que as distribuições condicionais dos parâmetros da distribuição normal associada ao *cluster* ϕ_j, como em (5.26) e (5.27), são dadas por

$$P\left(\mu_j | \sigma_j^2, \mathbf{y}_{\phi_j}\right) = N\left(\frac{n_j \underset{\phi_j}{\bar{y}}}{n_j + 0,01}, \frac{\sigma_j^2}{n_j + 0,01}\right)$$

e

$$P\left(\sigma_j^2 | \mathbf{y}_{\phi_j}\right) = IG\left(\frac{3 + n_j}{2}, \frac{2\left(n_j + 0,01\right) + \left(n_j + 0,01\right)\left(n_j - 1\right) \underset{\phi_j}{s^2} + n_j(0,01) \underset{\phi_j}{\bar{y}^2}}{2\left(n_j + 0,01\right)}\right)$$

para $j = 1, 2, ..., 6$.

Definidas as probabilidades condicionais de $c_i | c_{-i}, y_i$ e a distribuição *a posteriori* $\mu_j, \sigma_j^2 | \mathbf{c}, \phi_j, \mathbf{y}_{\phi_j}$, aplicamos o método *Gibbs sampling* como descrito em (5.28).

Para análise de convergência geramos duas cadeias, cada uma com $P = 20.000$ amostras das quais as primeiras 10.000 foram descartadas e consideramos um salto de 5. Com isso obtemos para cada uma das cadeias uma amostra de tamanho 2.000. Para cada cadeia calculamos a proporção de vezes que cada observação y_i foi associada a cada um dos n *clusters* e associamos cada y_i ao *cluster* que possuir maior proporção de associ-

amento. As Tabelas 23 e 24, mostram as proporções de associamento de cada observação y_i com cada um dos n *clusters*, respectivamente, para a cadeia 1 e cadeia 2, $i = 1, 2, ..., 5$.

Nas duas cadeias todas as observações foram identificadas como pertencentes a um mesmo *cluster* (*cluster* 2 na cadeia 1 e *cluster* 3 na cadeia 2, ver a proporções em destaque). Os resultados apresentados na Tabela 25 foram baseados nos valores gerados para os parâmetros da distribuição normal associada as observações destes dois *cluster*, ou seja, a média e o intervalo de credibilidade foram calculados baseados nos 4000 valores (2000 dos parâmetros da distribuição normal do *cluster* 2 da cadeia 1 + 2.000 dos parâmetros da distribuição normal do *cluster* 3 da cadeia 2). O verdadeiro valor da média e da variância da qual as observações foram geradas, $N(2, 1)$, está contido nos intervalos de credibilidade de 95%. Para verificar a convergência dos valores gerados para os parâmetros μ e σ^2 da distribuição normal associada aos *clusters*, utilizamos o diagnóstico de Gelmam e Rubin disponível no recurso CODA. Como os cinco valores foram detectados como pertencentes a um mesmo *cluster* (nas duas cadeias geradas), então não temos evidências de que as cinco observações pertençam a distribuições de probabilidades diferentes.

Tabela 23: Proporção de associamento, cadeia 1.

Observações	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
y_1	0,1709	**0,2260**	0,2198	0,1593	0,2240
y_2	0,1638	**0,2298**	0,2203	0,1606	0,2263
y_3	0,1663	**0,2296**	0,2207	0,1595	0,2238
y_4	0,1601	**0,2296**	0,2220	0,1645	0,2241
y_5	0,1635	**0,2296**	0,2210	0,1596	0,2263

Tabela 24: Proporção de associamento, cadeia 2.

Observações	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
y_1	0,1595	0,2034	**0,2366**	0,1937	0,2067
y_2	0,1543	0,2076	**0,2371**	0,1946	0,2064
y_3	0,1540	0,2026	**0,2386**	0,1947	0,2099
y_4	0,1533	0,2053	**0,2349**	0,1978	0,2086
y_5	0,1531	0,2042	**0,2369**	0,1974	0,2084

Tabela 25: Valores obtidos dos valores gerados para os parâmetros da distribuição associada aos clusters 2 e 3.

Parâmetros	Média	Int. de Cred. (95%)	Diag. Gelmam e Rubin
μ	2,0126	(-0,2805 ; 4,1069)	1
σ^2	1,0644	(0,1942 ; 3,9651)	1

5.3 Aplicação: Análise da expressão gênica

Utilizando a modelagem (5.4) para análise da expressão gênica, temos um modelo com mistura infinita equivalente ao modelo de misturas de processos Dirichlet que "aglomera" níveis de expressão gênica. Com isso obtemos *clusters* de genes que apresentam níveis de expressão similares.

Aqui voltamos para a situação de uma condição de tratamento, como descrito nos Capítulos 2 e 3.

Assim, para cada gene g em estudo, $g = 1, 2, ..., G$, temos um conjunto de variáveis observáveis $x^c_{g1}, ..., x^c_{gn_c}$ e $x^t_{g1}, ..., x^t_{gn_t}$, independentes, representando o logaritmo dos níveis de expressão do gene g na situação controle (c) e tratamento (t), onde

$$\begin{aligned} x^c_{g1}, ..., x^c_{gn_c} &\sim N\left(\mu_{gc}, \sigma^2_{gc}\right) \\ x^t_{g1}, ..., x^t_{gn_t} &\sim N\left(\mu_{gt}, \sigma^2_{gt}\right) \end{aligned}$$

com $\bar{x}_{gc}$ e $\bar{x}_{gt}$ sendo a média amostral das observações de controle (c) e de tratamento (t), respectivamente, onde

$$\begin{aligned} \bar{x}_{gc} &\sim N\left(\mu_{gc}, \frac{\sigma^2_{gc}}{n_{gc}}\right) \\ \bar{x}_{gt} &\sim N\left(\mu_{gt}, \frac{\sigma^2_{gt}}{n_{gt}}\right) \end{aligned}$$

com genes diferentes sendo tratados independentemente.

Considere τ_g como sendo o efeito de tratamento para o gene g, dado por

$$\tau_g = \mu_{gt} - \mu_{gc}. \tag{5.29}$$

Assim,

$$d_g = \bar{x}_t - \bar{x}_c, \tag{5.30}$$

que é a diferença entre a média amostral de tratamento e a média amostral de controle, que chamamos de estatística do efeito de tratamento. Logo, d_g é gerado segundo uma distribuição normal com média τ_g e variância σ_g^2,

$$d_g \sim N\left(\tau_g, \sigma_g^2\right), \tag{5.31}$$

onde

$$\sigma_g^2 = \frac{\sigma_{gc}^2}{n_c} + \frac{\sigma_{gt}^2}{n_t}.$$

Para análise da expressão gênica sobre o efeito de tratamento para cada um dos genes g, $g = 1, 2, ..., G$, utilizamos o modelo com mistura infinita em (5.4). Portanto, temos que

$$\begin{aligned} d_g|c, \phi, \tau_j, \sigma_j^2 &\sim N\left(\tau_j, \sigma_j^2\right) \\ Z_1, Z_2, ..., Z_G|w_1, w_2, ..., w_k &\sim Multinomial\left(1; w_1, w_2, ..., w_k\right) \\ \tau_j, \sigma_j^2 &\sim G_0 \\ w_1, w_2, ..., w_k &\sim Dirichlet\left(\frac{M}{k}, \frac{M}{k}, ..., \frac{M}{k}\right) \end{aligned} \tag{5.32}$$

com

$$M = a \text{ (constante)}, G_0\left(\tau_j|\sigma_j^2\right) = N\left(0, \frac{\sigma_j^2}{\lambda}\right), G_0\left(\sigma_g^2\right) = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right), \tag{5.33}$$

ou seja,

$$G_0\left(\tau_j, \sigma_j^2\right) = N\left(0, \frac{\sigma_j^2}{\lambda}\right) IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

com densidade

$$g_0\left(\tau_j, \sigma_j^2\right) = f_{N\left(0, \frac{\sigma_j^2}{\lambda}\right)}\left(\tau_g\right) f_{IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)}\left(\sigma_g^2\right)$$

para a, λ, α e β conhecidos e fixos e

$$c_g = \begin{cases} j, \text{ se a variável latente } Z_{ij} = 1 \\ 0, \text{ caso contrário} \end{cases},$$

para $g = 1, 2, .., G$ (Quantidade de genes em estudo), $j = 1, 2, ..., k$ e $k \to \infty$.

Logo, a probabilidade condicional para $c_g|c_{-g}, d_g$, como em (5.12), é dada por

$$\begin{aligned} P\left(c_g = j|c_{-g}\right) &= \frac{n_{-g,j}}{M+G-1} f_{N\left(\tau_j,\sigma_j^2\right)}\left(d_g\right) \\ P\left(c_g \neq j|c_{-g}\right) &= \frac{M}{M+G-1} q_{0i} P\left(\tau_j, \sigma_j^2|d_g\right) \end{aligned} \tag{5.34}$$

onde $f_{N\left(\tau_j,\sigma_j^2\right)}\left(d_g\right)$ é o valor da densidade da distribuição normal com média τ_j e variância σ_j^2 no ponto d_g, e

$$q_{og} = f_{t_\alpha\left(0,\frac{B(\lambda+1)}{\alpha\lambda}\right)}\left(d_g\right) \tag{5.35}$$

é o valor da densidade da distribuição *t-Student* com média zero, variância $\frac{B(\lambda+1)}{\alpha\lambda}$ e α graus de liberdade no ponto d_g, e

$$P\left(\tau_j, \sigma_j^2|d_g\right) = N\left(\frac{d_g}{\lambda+1}, \frac{\sigma_j^2}{\lambda+1}\right) IG\left(\frac{\alpha+2}{2}, \frac{\beta\left(\lambda+1\right)+\lambda d_g^2}{2\left(\lambda+1\right)}\right) \tag{5.36}$$

com distribuições condicionais dadas por

$$P\left(\tau_g|\sigma_g^2, d_g\right) = N\left(\frac{d_g}{\lambda+1}, \frac{\sigma_j^2}{\lambda+1}\right) \tag{5.37}$$

$$P\left(\sigma_j^2|d_g\right) = IG\left(\frac{\alpha+2}{2}, \frac{\beta\left(\lambda+1\right)+\lambda d_g^2}{2\left(\lambda+1\right)}\right), \tag{5.38}$$

onde $N\left(\frac{d_g}{\lambda+1}, \frac{\sigma_j^2}{\lambda+1}\right)$ e $IG\left(\frac{\alpha+2}{2}, \frac{\beta(\lambda+1)+\lambda d_g^2}{2(\lambda+1)}\right)$ representam, respectivamente, a distribuição normal com média $\frac{d_g}{\lambda+1}$ e variância $\frac{\sigma_j^2}{\lambda+1}$ e a distribuição gama inversa com parâmetro de forma $\frac{\alpha+2}{2}$ e parâmetro de escala $\frac{\beta(\lambda+1)+\lambda d_g^2}{2(\lambda+1)}$ (os cálculos de q_{oi} e $P\left(\tau_j, \sigma_j^2|d_g\right)$ é semelhante aos apresentados no caso 2: variâncias desconhecidas, na subseção 4.4.1 do Capítulo 4), para $g = 1, 2, ...G$ e $j = 1, 2, ..., k^*$.

A constante normalizadora apropriada, como em (5.15), é dada por

$$b = \left(\frac{n_{-g,j} f_{N\left(\tau_j,\sigma_j^2\right)}\left(d_g\right) + M f_{t_\alpha\left(0,\frac{B(\lambda+1)}{\alpha\lambda}\right)}\left(d_g\right)}{M+G-1}\right)^{-1}.$$

para $g = 1, 2, ..., G$ e $j = 1, 2, ..., k^*$.

Com sugerido no algoritmo 2.1, fazendo $k^* = G$ (quantidade de genes em estudo),

dado os valores de $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ definidas como em (5.34), temos as quantidades $n_1, n_2, ..., n_G$ pertencentes aos *clusters* $\phi_1, \phi_2, ..., \phi_G$, respectivamente. Para cada *cluster* ϕ_j, como em (5.24), temos uma função de verossimilhança associada, que é dada por,

$$L\left(\tau_j, \sigma_j^2 | \mathbf{d}_{\phi_j}\right) \propto \left(\sigma_j^2\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\sigma_j^2}\left[\left(\underset{\phi_j}{\bar{d}} - \delta_j\right)^2 + (n_j - 1)\, \underset{\phi_j}{s^2}\right]\right\} \tag{5.39}$$

onde onde $\mathbf{d}_{\phi_j} = \begin{pmatrix} d_1, d_2, ..., d_{n_j} \\ \phi_j \quad \phi_j \qquad \phi_j \end{pmatrix}$ representa as n_j medidas pertencentes ao *cluster* ϕ_j e $\underset{\phi_j}{\bar{d}}$ e $\underset{\phi_j}{s^2}$ é a média e a variância amostral das medidas pertencentes a este *cluster*, para $g = 1, 2, ..., G$ e $j = 1, 2, ..., G$.

Logo, como em (5.25) a distribuição *a posteriori* dos parâmetros τ_j e σ_j^2, da distribuição normal associada ao *cluster* ϕ_j dado $\mathbf{c} = (c_1, c_2, ..., c_n)$ e as observações $\mathbf{d}_{\phi_j}$, é dada por

$$\begin{aligned} P\left(\tau_j, \sigma_j^2 | \mathbf{d}_{\phi_j}\right) &= N\left(\frac{n_j \underset{\phi_j}{\bar{d}}}{n_j + \lambda}, \frac{\sigma_j^2}{n_j + \lambda}\right) \\ & \quad IG\left(\frac{\alpha + n_j + 1}{2}, \frac{\beta\,(n_j + \lambda) + (n_j + \lambda)\,(n_j - 1)\, \underset{\phi_j}{s^2} + n_j \lambda \underset{\phi_j}{\bar{d}^2}}{2\,(n_j + \lambda)}\right), \end{aligned} \tag{5.40}$$

para $g = 1, 2, ..., G$ e $j = 1, 2, ..., G$.

As distribuições condicionais, como em (5.26) e (5.27), utilizadas no método *Gibbs Samplig* para estimação dos parâmetros, são dadas por

$$P\left(\tau_j | \sigma_j^2, \mathbf{y}_{\phi_j}\right) = N\left(\frac{n_j \underset{\phi_j}{\bar{d}}}{n_j + \lambda}, \frac{\sigma_j^2}{n_j + \lambda}\right) \tag{5.41}$$

$$P\left(\sigma_j^2 | \mathbf{d}_{\phi_j}\right) = IG\left(\frac{\alpha + n_j + 1}{2}, \frac{\beta\,(n_j + \lambda) + (n_j + \lambda)\,(n_j - 1)\, \underset{\phi_j}{s^2} + n_j \lambda \underset{\phi_j}{\bar{d}^2}}{2\,(n_j + \lambda)}\right). \tag{5.42}$$

Definido as probabilidades condicionais de $c_i | c_{-i}, y_i$, e a distribuição *a posteriori* de $\tau_j, \sigma_j^2 | \mathbf{d}_{\phi_j}$ inicializamos o método *Gibbs sampling* fazendo $c_i^0 = i$, ou seja, inicialmente cada observação em um *cluster*, $\tau_g = d_g$ e $\sigma_g^2 = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$, para $g = 1, 2, ..., G$ e a *p-ésima*

sequência gerada pelo método *Gibbs sampling,* $p = 1, 2, ..., P$, e a determinação do *cluster* associado a cada d_g, para $g = 1, 2, ..., G$, é feita como em (5.28).

Terminadas as P seqüências geradas pelo método *Gibbs sampling*, temos formados k' *clusters* de níveis de expressão similares devido ao efeito de tratamento, onde k' representa a quantidade de *clusters*, dentre os G possíveis, que possuem pelo menos uma observação. Para determinarmos quais dos k' *clusters* de genes apresentam evidências para níveis de expressão diferentes entre controle e tratamento, supomos que:

- se o *cluster* ϕ_j, $j = 1, 2, ..., k'$, é composto por medidas de níveis de expressão que não apresentam evidências para níveis de expressão diferentes entre a situação de controle e a situação de tratamento, então as medidas d_g dos genes g que são associados a este *cluster*, são gerados de uma distribuição normal com média zero e variância $\underset{M_0}{\sigma^2_{\phi_j}}$. Esta condição chamamos de modelo M_0.

- se o *cluster* ϕ_j, $j = 1, 2, ..., k'$, é composto por medidas de níveis de expressão que apresentam evidências para níveis de expressão diferentes entre a situação de controle e a situação de tratamento, então as medidas d_g dos genes g que são associados a este *cluster*, são gerados de uma distribuição normal com média μ_{ϕ_j} (para $\mu_{\phi_j} \neq 0$) e variância $\underset{M_1}{\sigma^2_{\phi_j}}$. Esta condição chamamos de modelo M_1.

Para cada um dos modelos temos as respectivas verossimilhanças, dadas por

$$\underset{\phi_j}{L_{M_0}}\left(\underset{M_0}{\sigma^2_{\phi_j}}|\mathbf{d}_{\phi_j}\right) \propto \left(\underset{M_0}{\sigma^2_{\phi_j}}\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\underset{M_0}{\sigma^2_{\phi_j}}}\sum_{i=1}^{n_j}\underset{\phi_j}{d_i^2}\right\} \tag{5.43}$$

$$\begin{aligned}\underset{\phi_j}{L_{M_1}}\left(\mu_{\phi_j}, \underset{M_1}{\sigma^2_{\phi_j}}|\mathbf{d}_{\phi_j}\right) &\propto \left(\underset{M_1}{\sigma^2_{\phi_j}}\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\underset{M_1}{\sigma^2_{\phi_j}}}\sum_{i=1}^{n_j}\left(\underset{\phi_j}{d_i} - \mu_{\phi_j}\right)^2\right\} \\ &\propto \left(\underset{M_1}{\sigma^2_{\phi_j}}\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{n_j\left(\underset{\phi_j}{\bar{d}} - \mu_{\phi_j}\right)^2 + (n-1)\,\underset{\phi_j}{s^2}}{2\underset{M_1}{\sigma^2_{\phi_j}}}\right\}\end{aligned} \tag{5.44}$$

para $j = 1, 2, ..., k'$.

Considerar se há ou não evidências para níveis de expressão diferentes para as medidas

de um determinado *cluster* ϕ_j, equivale a considerar o modelo M_0 ou M_1 como sendo o modelo que melhor explica as medidas de níveis de expressão d_g pertencentes ao *cluster* ϕ_j, para $g = 1, 2, ..., G$ e $j = 1, 2, ..., k$'. Dessa forma, consideramos a abordagem bayesiana e o critério DIC (como descrito abaixo), para selecionar qual o modelo mais adequado e assim determinar se o *cluster* é composto por medidas de níveis de expressão diferentes ou não.

Optamos por utilizar o critério DIC e não o fator de Bayes devido ao cálculo da aproximação do fator de Bayes ser inviável para algumas situações, tais como, quando uma grande quantidade de genes com variância pequena é associada a um mesmo *cluster*. O cálculo do valor da função de verossimilhança, associado a este *cluster*, se torna indeterminado computacionalmente, devido a variância ser pequena e estar elevada a um expoente "grande".

5.3.1 Abordagem bayesiana e DIC

Para a abordagem Bayesiana, devemos especificar distribuições *a priori* para os parâmetros $\sigma^2_{\substack{\phi_j \\ M_0}}$ μ_{ϕ_j} e $\sigma^2_{\substack{M_1 \\ \phi_j}}$, para $j = 1, 2, ..., k$', dos modelos M_0 e M_1, respectivamente, e fazemos inferências baseados em suas distribuições *a posteriori*.

Consideramos o conjunto de hiperparâmetros $\varpi = (\lambda, \alpha, \beta)$ e as distribuições *a priori* usuais, normal e gama inversa dadas por

$$\pi\left(\sigma^2_{\substack{\phi_j \\ M_0}}\right) = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

ou seja

$$\pi\left(\sigma^2_{\substack{\phi_j \\ M_0}}\right) \propto \left(\sigma^2_{\substack{\phi_j \\ M_0}}\right)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_0}}}\right\}$$

e

$$\pi\left(\mu_{\phi_j} | \sigma^2_{\substack{\phi_j \\ M_1}}\right) = N\left(0, \frac{\sigma^2_{\substack{M_1 \\ \phi_j}}}{\lambda}\right)$$

ou seja,

$$\pi\left(\underset{M_1}{\mu_{\phi_j}}|\underset{M_1}{\sigma^2_{\phi_j}}\right) \propto \left(\underset{M_1}{\sigma^2_{\phi_j}}\right)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda}{2\underset{M_1}{\sigma^2_{\phi_j}}}\mu^2_{\phi_j}\right\}$$

e

$$\pi\left(\underset{M_1}{\sigma^2_{\phi_j}}\right) = IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right)$$

ou seja,

$$\pi\left(\underset{M_1}{\sigma^2_{\phi_j}}\right) \propto \left(\underset{M_1}{\sigma^2_{\phi_j}}\right)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left\{-\frac{\beta}{2\underset{M_1}{\sigma^2_{\phi_j}}}\right\},$$

para $j = 1, 2, ..., k$.

Aplicando o teorema de Bayes, temos que a distribuição *a posteriori* para o modelo M_0 é dada por

$$\begin{aligned}
\pi\left(\underset{M_0}{\sigma^2_{\phi_j}}|\mathbf{d}_{\phi_j}\right) &\propto L_{M_0}\left(\underset{M_0}{\sigma^2_{\phi_j}}|\mathbf{d}_{\phi_j}\right)\pi\left(\underset{M_0}{\sigma^2_{\phi_j}}\right) \qquad (5.45)\\
&\propto \left(\underset{M_0}{\sigma^2_{\phi_j}}\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\underset{M_0}{\sigma^2_{\phi_j}}}\sum_{i=1}^{n_j} \underset{\phi_j}{d_i^2}\right\}\left(\underset{M_0}{\sigma^2_{\phi_j}}\right)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left\{-\frac{\beta}{2\underset{M_0}{\sigma^2_{\phi_j}}}\right\}\\
&\propto \left(\underset{M_0}{\sigma^2_{\phi_j}}\right)^{-\left(\frac{\alpha+n_j}{2}+1\right)} \exp\left\{-\frac{\beta}{2\underset{M_0}{\sigma^2_{\phi_j}}} - \frac{1}{2\underset{M_0}{\sigma^2_{\phi_j}}}\sum_{i=1}^{n_j} \underset{\phi_j}{d_i^2}\right\}\\
&\propto \left(\underset{M_0}{\sigma^2_{\phi_j}}\right)^{-\left(\frac{\alpha+n_j}{2}+1\right)} \exp\left\{-\frac{1}{\underset{M_0}{\sigma^2_{\phi_j}}}\left(\frac{\beta + \sum_{i=1}^{n_j} \underset{\phi_j}{d_i^2}}{2}\right)\right\}
\end{aligned}$$

A distribuição *a posteriori* para o modelo M_1 é dada por

$$\pi\left(\underset{M_1}{\mu_{\phi_j}}, \underset{M_1}{\sigma^2_{\phi_j}}|\mathbf{d}_{\phi_j}\right) \propto \underset{\phi_j}{L_{M_1}}\left(\underset{M_1}{\mu_{\phi_j}}, \underset{M_1}{\sigma^2_{\phi_j}}|\mathbf{d}_{\phi_j}\right)\pi\left(\mu_{\phi_j}\right)\pi\left(\underset{M_1}{\sigma^2_{\phi_j}}\right) \qquad (5.46)$$

$$
\begin{aligned}
&\propto \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{n_j\left(\bar{d}_{\phi_j} - \mu_{\phi_j}\right)^2 + (n-1)\, s^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\right\} \\
&\quad \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\mu^2_{\phi_j}\right\} \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\right\} \\
&\propto \exp\left\{-\frac{n_j\left(\bar{d}_{\phi_j} - \mu_{\phi_j}\right)^2 + (n-1)\, s^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{\lambda}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\mu^2_{\phi_j}\right\} \\
&\quad \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\right\} \\
&\propto \exp\left\{-\frac{n_j\left(\bar{d}_{\phi_j} - \mu_{\phi_j}\right)^2}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{\lambda}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\mu^2_{\phi_j}\right\} \\
&\quad \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{(n-1)\, s^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\right\} \\
&\propto \exp\left\{-\frac{1}{\sigma^2_{\substack{\phi_j \\ M_1}}}\left(n_j \bar{d}^2_{\phi_j} - 2\mu_j n_j \bar{d}_{\phi_j} + n_j\mu_{\phi_j} + \lambda\mu^2_{\phi_j}\right)\right\} \\
&\quad \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{(n-1)\, s^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\left((n_j+\lambda)\,\mu^2_{\phi_j} - 2\mu_{\phi_j} n_j \bar{d}_{\phi_j}\right)\right\} \\
&\quad \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{(n-1)\, s^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{n_j \bar{d}^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\right\} \\
&\propto \exp\left\{-\frac{n_j+\lambda}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\left(\mu^2_j - 2\mu_j\left(\frac{n_j \bar{d}_{\phi_j}}{n_j+\lambda}\right)\right)\right\}
\end{aligned}
$$

$$
\begin{aligned}
& \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{(n-1)\,s^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{n_j\,\bar{d}^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}}\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2\frac{\sigma^2_{\substack{\phi_j \\ M_1}}}{n_j+\lambda}}\left(\mu^2_{\phi_j} - \frac{n_j\,\bar{d}_{\phi_j}}{n_j+\lambda}\right)^2\right\} \\
& \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{\beta}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{(n-1)\,s^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}} - \frac{n_j\,\bar{d}^2_{\phi_j}}{2\sigma^2_{\substack{\phi_j \\ M_1}}} + \frac{\left(n_j\,\bar{d}_{\phi_j}\right)^2}{2\sigma^2_{\substack{\phi_j \\ M_1}}(n_j+\lambda)}\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2\frac{\sigma^2_{\substack{\phi_j \\ M_1}}}{n_j+\lambda}}\left(\mu^2_{\phi_j} - \frac{n_j\,\bar{d}_{\phi_j}}{n_j+\lambda}\right)^2\right\} \\
& \left(\sigma^2_{\substack{\phi_j \\ M_1}}\right)^{-\left(\frac{\alpha+n_j+1}{2}+1\right)} \exp\left\{-\frac{1}{\sigma^2_{\substack{\phi_j \\ M_1}}}\left(\frac{\beta\,(n_j+\lambda) + (n_j+\lambda)\,(n-1)\,s^2_{\phi_j} + n_j\lambda\bar{d}^2_{\phi_j}}{2}\right)\right\} \\
= \quad & N\left(\frac{n_j\,\bar{d}_{\phi_j}}{n_j+\lambda}, \frac{\sigma^2_{\substack{\phi_j \\ M_1}}}{n_j+\lambda}\right) IG\left(\frac{\alpha+n_j+1}{2}, \frac{\beta\,(n_j+\lambda) + (n_j+\lambda)\,(n-1)\,s^2_{\phi_j} + n_j\lambda\bar{d}^2_{\phi_j}}{2}\right)
\end{aligned}
$$

para $j = 1, 2, ..., k$.

Para análise da expressão gênica, selecionamos o modelo M_0 ou M_1, para um determinado *cluster* ϕ_j, $j = 1, 2, ..., k$, utilizamos o DIC.

Para cada um dos modelos, M_0 e M_1 e um determinado *cluster* ϕ_j, temos que as *deviances*, como em (3.14), são dadas por

$$
\begin{aligned}
D_{\substack{M_0 \\ \phi_j}}\left(\sigma^2_{\substack{M_1 \\ \phi_j}}\right) &= -2\log L_{M_0}\left(\sigma^2_{\substack{\phi_j \\ M_0}} | \mathbf{d}_{\phi_j}\right) \\
&\propto -2\log\left[\left(\sigma^2_{\substack{\phi_j \\ M_0}}\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2\sigma^2_{\substack{\phi_j \\ M_0}}}\sum_{i=1}^{n_j} d^2_{i\,\phi_j}\right\}\right] \\
&\propto n_j \log\left(\sigma^2_{\substack{\phi_j \\ M_0}}\right) + \frac{\sum_{i=1}^{n_j} d^2_{i\,\phi_j}}{\sigma^2_{\phi_j}}
\end{aligned}
\tag{5.47}
$$

$$
\begin{aligned}
\underset{\phi_j}{D_{M_1}}\left(\underset{M_1}{\mu_{\phi_j}, \sigma^2_{\phi_j}}\right) &= -2\log L_{M_1}\left(\underset{M_1}{\mu_{\phi_j}, \sigma^2_{\phi_j}} | \mathbf{d}_{\phi_j}\right) \qquad (5.48)\\
&\propto -2\log\left[\left(\underset{M_1}{\sigma^2_{\phi_j}}\right)^{-\frac{n_j}{2}} \exp\left\{-\frac{n_j\left(\underset{\phi_j}{\bar{y}} - \mu_{\phi_j}\right)^2 + (n-1)\underset{\phi_j}{s^2}}{2\underset{M_1}{\sigma^2_{\phi_j}}}\right\}\right]\\
&\propto n_j \log\left(\underset{M_1}{\sigma^2_{\phi_j}}\right) + -\frac{n_j\left(\underset{\phi_j}{\bar{d}} - \mu_{\phi_j}\right)^2 + (n-1)\underset{\phi_j}{s^2}}{\underset{M_1}{\sigma^2_{\phi_j}}}
\end{aligned}
$$

para $j = 1, 2, ..., k$.

O valor DIC associado aos modelos M_0 e M_1, como em (3.15), são dados por

$$\underset{\phi_j}{\text{DIC}_{M_0}} = \underset{\phi_j}{\bar{D}_{M_0}} + \underset{\phi_j}{P_{D_{M_0}}}$$

$$\underset{\phi_j}{\text{DIC}_{M_1}} = \underset{\phi_j}{\bar{D}_{M_1}} + \underset{\phi_j}{P_{D_{M_1}}}$$

onde

$$
\begin{aligned}
\underset{\phi_j}{P_{D_{M_0}}} &= \underset{\phi_j}{\bar{D}_{M_0}} - D\left(\underset{M_0}{\bar{\sigma}^2_{\phi_j}}\right)\\
\underset{\phi_j}{P_{D_{M1}}} &= \underset{\phi_j}{\bar{D}_{M_1}} - D\left(\bar{\mu}_j, \underset{M_1}{\sigma^2_{\phi_j}}\right)
\end{aligned}
$$

com $\underset{M_0}{\bar{\sigma}^2_{\phi_j}}$ sendo a média dos valores gerados para $\underset{M_0}{\sigma^2_{\phi_j}}$ de sua distribuição *a posteriori* (5.45) e $\bar{\mu}_{\phi_j}, \underset{M_1}{\bar{\sigma}^2_{\phi_j}}$ sendo a média dos valores gerados para μ_{ϕ_j},$\underset{M_1}{\sigma^2_{\phi_j}}$ de sua distribuição *a posteriori* (5.46), para $j = 1, 2, ..., k$.

O cálculo do valor DIC para os modelos M_0 e M_1, segue os mesmos passos descritos na seção 3.4.

5.3.2 Simulação

Desenvolvemos para o modelo com mistura infinita e para o DIC, o mesmo estudo de simulação realizado para os métodos estatísticos anteriores, com objetivo de verificar seu

comportamento na detecção de *clusters* com e sem evidências para níveis de expressão diferentes, quando consideramos diferentes afastamentos na média e na variância (ou ambos) das observações de tratamento com relação as observações de controle. Porém, agora, utilizamos 100 genes (simulados) ao invez de 1000 genes (simulados) como nos métodos anteriores. Isto foi feito devido ao tempo de simulação com 1000 genes ser muito "alto".

Os hiperparâmetros utilizados para a formação dos *clusters* foram os mesmos da simulação realizada para o modelo MPD. Isto é, $M = 1$, $\lambda = 0,01$, $\alpha = 2$, $\beta = 2$, $\beta = 1$ e $\beta = 0,5$.

Os diferentes valores para β foram utilizados para verificarmos a sensibilidade do modelo com mistura infinita na detecção dos *clusters* com relação a escolha dos hiperparâmetros.

Para cada valor de β, temos as probabilidades condicionais dadas em (5.34) e as distribuições condicionais dos *clusters* com pelo menos uma observação ($n_{-i,j} \geq 1$) dadas em (5.41) e (5.42).

Para aplicação do método *Gibbs sampling*, utilizamos $P = 1000$ iterações. Para definirmos os *clusters*, calculamos a proporção de vezes que cada medida d_g foi associada ao *cluster* ϕ_j durante as $P = 1000$ iterações realizadas e definimos os *clusters* como descrito em (5.28).

Definidos os *clusters*, aplicamos o critério DIC como descrito na subseção 5.3.1 utilizando $m = 1000$, para detectar os *cluster* que apresentam evidências para diferença. Os hiperparâmetros utilizados no cálculo do DIC foram os mesmo utilizados na formação dos *cluster*s. Por exemplo, se utilizamos $\beta = 2$ para formar os *cluster*s também utilizamos $\beta = 2$ para calcular o valor DIC associado aos modelos, M_0 e M_1, de cada *cluster* formado.

As Tabelas 26, 27 e 28 mostram as quantidades de genes (simulados) detectados com evidências para diferença e a quantidade de *cluster*s formados (indicados pela letra **C** à direita de cada valor de δ) para cada valor de δ e γ utilizados.

Cada linha mostra as quantidades de genes detectados com evidências para diferença e a quantidade de *cluster*s formados para γ fixo e δ variando. Cada coluna mostra a quantidade de genes detectados com evidências para diferença e a quantidade de *cluster*s formados para δ fixo e γ variando.

Afastando-se da média de controle ($\delta \neq 0$) e para $\gamma = 0,25$ e $\gamma = 1$, o modelo com mistura infita e o DIC detectam todos os genes (simulados) como pertencentes a um mesmo *cluster* que apresenta evidências para diferença (ver Tabelas 26, 27 e 28).

Esta característica já era esperada, pois como as observações de tratamento são geradas de uma distribuição normal com média $\mu_{gt} = \mu_{gc} + \delta$ e variância $\sigma^2_{gt} = (\gamma\sigma_{gc})^2$então $\tau_g = \delta$ e devido a "pequena" variabilidade de controle $\sigma^2_c = 0,04$, todas as medidas d_g são concentradas em torno de δ, para $g = 1, 2, ...100$. Logo, temos somente um *cluster* com evidências para diferença devido suposição do modelo M_0 no critério DIC, que considera que d_g foi gerada de um distribuição normal com média zero e variância $\sigma^2_{\underset{M_0}{\phi_j}}$, para $j = 1, 2, ..., 100$. Por exemplo, quando utilizamos $\delta = 0,5$ e $\gamma = 1$, temos formado apenas um *cluster*, indicado por $\mathbf{C}_1$, com evidências para diferença. Como pode ser observado na Figura 16. Para $\beta = 2$, $\beta = 1$ e $\beta = 0,5$ os resultados são os mesmos.

Para δ fixo, Tabelas 26, 27 e 28, a medida que aumentamos o valor de γ, isto é, aumentamos a variabilidade das observações de tratamento com relação as observações de controle, o modelo com misturas infinita e o DIC detectam uma quantidade maior de *cluster*s. O que é positivo, pois como temos medidas d_g mais dispersas devido aos valores da condição de tratamento serem gerados de uma distribuição normal com uma maior variabilidade, temos uma maior heterogeneidade.

Por exemplo, para $\delta = 0$ e $\gamma = 16$ temos 3 *cluster* formados, que indicamos por $\mathbf{C}_1$, $\mathbf{C}_2$ e $\mathbf{C}_3$, onde:

- $\mathbf{C}_1$: composto por medidas d_g somente positivas;
- $\mathbf{C}_2$: composto por medidas d_g positivas e negativas;
- $\mathbf{C}_3$: composto por medidas d_g somente negativas;

com $\mathbf{C}_1$ e $\mathbf{C}_3$ sendo detectados com evidências para diferença . Ou seja, para $\delta = 0$ aumentando o valor de γ temos uma quantidade maior de medidas d_g distante da reta $y = x$. Isto justifica o fato de aumentarmos o valor de γ, aumentar a quantidade de *cluster*s formados e aumentar a quantidade de genes detectados com evidências para diferença do que a variação γ anterior (ver Figuras 17, 18 e 19).

Para $\delta = -0,5$ e $\gamma = 9$ também temos três *cluster*s $\mathbf{C}_1$, $\mathbf{C}_2$ e $\mathbf{C}_3$, onde:

- $\mathbf{C}_1$: composto por medidas d_g somente positivas;
- $\mathbf{C}_2$: composto por medidas d_g positivas e negativas;
- $\mathbf{C}_3$: composto por medidas d_g somente negativas;

com $\mathbf{C}_1$ e $\mathbf{C}_3$ sendo detectados com evidências para diferença. Ou seja, para $\delta = -0,5$ aumentando o valor de γ temos uma quantidade maior de medidas d_g distante de $\delta = -0,5$ e próximos a reta $y = x$. Isto justifica o fato de aumentarmos o valor de γ, aumentar a quantidade de *cluster*s formados e diminuir a quantidade de genes detectados com evidências para diferença do que a variação γ anterior.

Os resultados obtidos com $\beta = 2$, $\beta = 1$ e $\beta = 0,5$ são semelhantes, como pode ser observado pelas Tabelas 26, 27 e 28.

Para $\delta = \pm 0,5$ e $\delta = 0$, a quantidade de *cluster*s formados sempre é a mesma, com $\beta = 2$, $\beta = 1$ e $\beta = 0,5$, porém há uma diferença na quantidade de genes pertencentes a cada *cluster* e conseqüentemente há uma diferença na quantidade de genes detectados com evidências para diferença. Esta diferença é provocada pelos genes com medidas d_g na "fronteira" dos *cluster*s formados, pois estas medidas ficam alternando de *cluster*s dependendo da escolha do valor β. Isto é, para medidas d_g na "fronteira" dos *cluster*s formados a escolha dos hiperparâmetros das distribuições *a priori* pode influenciar o gene a pertencer a um determinado *cluster* e conseqüentemente ser detectado ou não com evidências para diferença. Por exemplo, podemos notar pelas Figuras 17, 18 e 19, que mostram os *clusters* formados com $\beta = 2$, $\beta = 1$ e $\beta = 0,5$, que alguns genes na "fronteira"dos *cluster*s formados mudam de *cluster* ao mudarmos o valor de β.

Para $\delta = \pm 1$ e $\delta = \pm 0,8$ os resultados são iguais tanto em quantidade de *cluster*s formados quanto em quantidade de genes detectados com evidências para diferença. Porém, ocorre o mesmo problema com os genes com medidas d_g nas fronteiras dos *cluster*s formados.

Portanto, temos que o modelo com mistura infinita é sensível a escolha dos hiperparâmetros levando a resultados diferentes. Assim, devemos ter certa atenção para definirmos os valores dos hiperparâmetros, se possível nos basear na opinião de um especialista ou

utilizar métodos empíricos. Pois assim estaremos trabalhando com informação *a priori* o que possivelmente minimizará esta mudança de *cluster*s pelas medidas d_g na "fronteira"dos *clusters* formados.

Tabela 26: Quantidades detectadas com evidência para diferença e quantidades de *cluster*s, com β=2.

	δ													
γ	-1,0	C	-0,8	C	-0,5	C	0,0	C	0,5	C	0,8	C	1,0	C
0,25	100	1	100	1	100	1	0	1	100	1	100	1	100	1
1	100	1	100	1	100	1	0	1	100	1	100	1	100	1
4	100	2	100	2	100	2	09	3	100	2	100	2	100	2
9	100	3	100	3	90	3	45	3	93	3	100	3	100	3
16	100	3	100	3	86	3	47	3	68	3	100	3	100	3

Tabela 27: Quantidades detectadas com evidência para diferença e quantidades de *cluster*s, com β=1.

	δ													
γ	-1,0	C	-0,8	C	-0,5	C	0,0	C	0,5	C	0,8	C	1,0	C
0,25	100	1	100	1	100	1	0	1	100	1	100	1	100	1
1	100	1	100	1	100	1	0	1	100	1	100	1	100	1
4	100	2	100	2	100	2	09	3	100	2	100	2	100	2
9	100	3	100	3	90	3	45	3	100	3	100	3	100	3
16	100	3	100	3	86	3	47	3	72	3	100	3	100	3

Tabela 28: Quantidades detectadas com evidência para diferença e quantidades de *cluster*s, com β=0,5.

	δ													
γ	-1,0	C	-0,8	C	-0,5	C	0,0	C	0,5	C	0,8	C	1,0	C
0,25	100	1	100	1	100	1	0	1	100	1	100	1	100	1
1	100	1	100	1	100	1	0	1	100	1	100	1	100	1
4	100	2	100	2	100	2	09	3	100	2	100	2	100	2
9	100	3	100	3	100	3	45	3	95	3	100	3	100	3
16	100	3	100	3	90	3	47	3	80	3	100	3	100	3

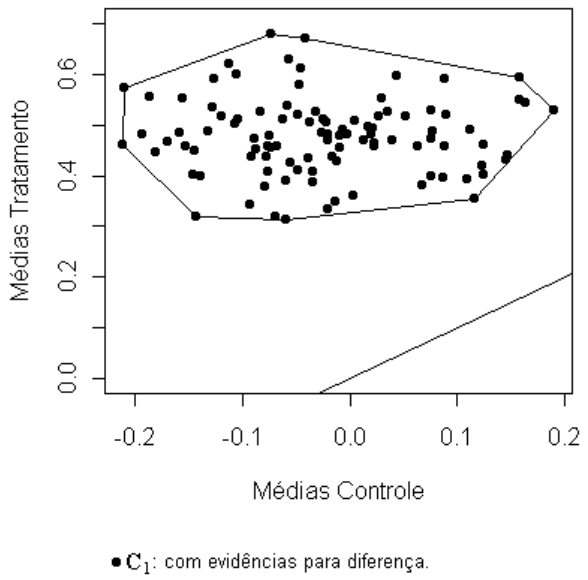

Figura 16: *cluster* $\mathbf{C}_1$, $\delta = 0,5$ e $\gamma = 1$.

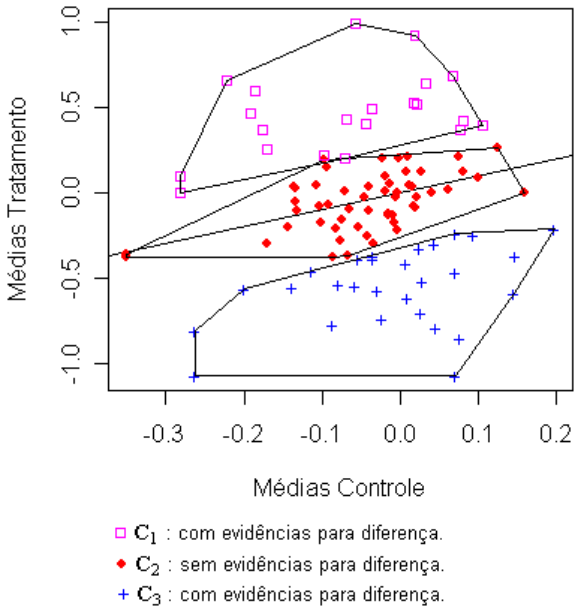

Figura 17: *cluster*s formados, $\delta = 0$, $\gamma = 16$ e $\beta = 2$.

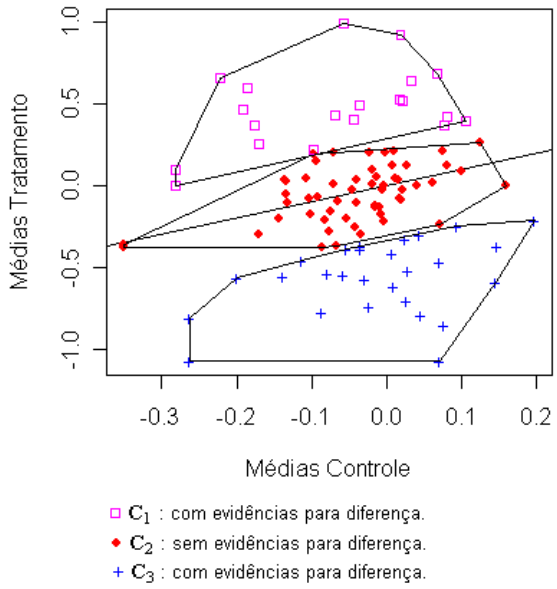

Figura 18: *cluster*s formados, $\delta = 0$, $\gamma = 16$ e $\beta = 1$.

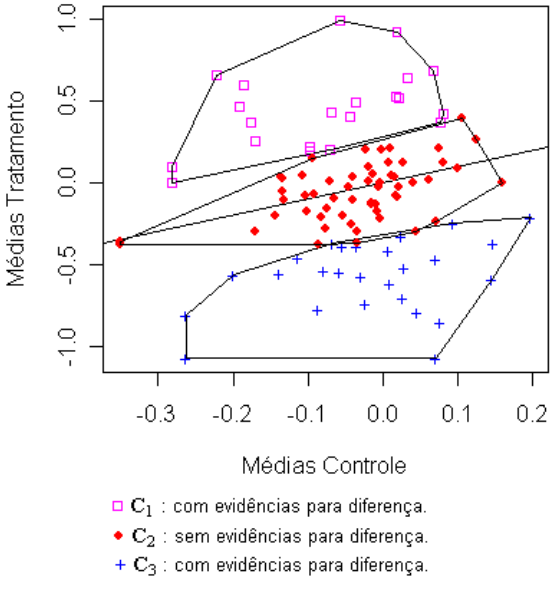

Figura 19: *cluster*s formados, $\delta = 0$, $\gamma = 16$ e $\beta = 0,5$

5.3.3 Aplicação

Aplicamos o modelo com mistura infinita equivalente ao modelo de misturas de processos Dirichlet (MPD), dado em (5.32) e o DIC no experimento realizado com células da bactéria *Escherichia Coli* com relação aos padrões IHF^+ e IHF^-, já utilizado nas

aplicações dos métodos estatísticos anteriores.

Como citado, dispomos de 434 genes, $g = 1, 2, ..., 434$, e uma condição de tratamento. Para cada um dos genes temos cinco medidas de níveis de expressão para a situação de controle (c) e cinco medidas de níveis de expressão para a situação de tratamento (t). Isto é,

- $x_1^c, ..., x_{n_c=5}^c$: representa as 5 medidas dos níveis de expressão para o g-ésimo gene sob a situação de controle (c);

- $x_1^t, ..., x_{n_t=5}^t$: representa as 5 medidas dos níveis de expressão para o g-ésimo gene sob a situação de tratamento (t);

- $\bar{x}_{gc}$: média dos níveis de expressão do g-ésimo gene na situação de controle (c);

- $\bar{x}_{gt}$: média dos níveis de expressão do g-ésimo gene na situação de tratamento (t);

- $d_g = \bar{x}_{gt} - \bar{x}_{gc}$: é a estatística do efeito de tratamento para o g-ésimo gene, para $g = 1, 2, ..., 434$.

Os hiperparâmetros utilizados na formação dos *clusters* e no cálculo DIC, foram os mesmos utilizados na simulação. Para cada valor de β, temos as probabilidades condicionais dadas em (5.34) e as distribuições condicionais dos *cluster*s com pelo menos uma observação, dadas como em (5.41) e (5.42).

Definidas as probabilidades condicionais e as distribuições condicionais dos *cluster*s, aplicamos o método *Gibbs sampling* como em (5.28) com $P = 5000$ iterações.

Formados os *cluster*, aplicamos o critério DIC para detectar quais *cluster* apresentam evidências para diferença, utilizando $m = 5000$ e os mesmo hiperparâmetros utilizados na formação dos *cluster*s.

As Figuras 20, 21 e 22 mostram as médias de controle e de tratamento para cada um dos genes g, $g = 1, 2, ..., 434$, destacando os *cluster*s formados e os *clusters* que apresentam evidências para diferença.

Para $\beta = 2$; distribuição *a priori* $IG(1, 1)$ para a variância, 4 *cluster*s foram formados, $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ e $\mathbf{C}_4$ onde:

- $\mathbf{C}_1$: é composto por 3 medidas d_g somente positivas;

- $\mathbf{C}_2$: é composto por 103 medidas d_g somente positivas;

- $\mathbf{C}_3$: é composto por 243 medidas d_g positivas e negativas;

- $\mathbf{C}_4$: é composto por 85 medidas d_g somente negativas;

com $\mathbf{C}_1$, $\mathbf{C}_2$ e $\mathbf{C}_4$ sendo detectados com evidências para diferença (ver Figura 20).

Para $\beta = 1$; distribuição *a priori* $IG(1; 0,5)$ para a variância, 4 *cluster*s foram formados, $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ e $\mathbf{C}_4$ onde:

- $\mathbf{C}_1$: é composto por 6 medidas d_g somente positivas;
- $\mathbf{C}_2$: é composto por 121 medidas d_g somente positivas;
- $\mathbf{C}_3$: é composto por 210 medidas d_g positivas e negativas;
- $\mathbf{C}_4$: é composto por 97 medidas d_g somente negativas;

com $\mathbf{C}_1$, $\mathbf{C}_2$ e $\mathbf{C}_4$ sendo detectados com evidências para diferença (ver Figura 21).

Para $\beta = 0,5$, distribuição *a priori* $IG(1; 0,25)$ para a variância, 4 *cluster*s foram formados, $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ e $\mathbf{C}_4$ onde:

- $\mathbf{C}_1$: é composto por 16 medidas d_g somente positivas;
- $\mathbf{C}_2$: é composto por 143 medidas d_g somente positivas;
- $\mathbf{C}_3$: é composto por 188 medidas d_g positivas e negativas;
- $\mathbf{C}_4$: é composto por 87 medidas d_g somente negativas;

com $\mathbf{C}_1$, $\mathbf{C}_2$ e $\mathbf{C}_4$ sendo detectados com evidências para diferença (ver Figura 22).

Comparando com resultados obtidos com as aplicações dos métodos anteriores, temos que:

- Todos os genes detectados com evidências para diferença pelo teste-t e pelo modelo de misturas de processos Dirichlet (MPD) também foram detectados pelo modelo com mistura infinita e DIC;

- Dos detectados pelo fator de Bayes e pelo critério DIC (Capítulo 3), somente os genes próximos a reta $y = x$, (genes 273, 323, 366 e 370 - Fator de Bayes e genes 273, 323, 370 - Critério DIC) não foram detectados pelo modelo com misturas infinita e DIC.

Os resultados obtidos com o modelo com mistura infinita e DIC para os genes com medidas d_g distantes da reta $y = x$ são satisfatórios, pois estes genes são detectados como

pertencentes a um *cluster* que apresenta evidências para diferença.

Porém, como observado na simulação, ocorrem problemas com os genes com medidas d_g na fronteira do *cluster* $\mathbf{C}_3$ com $\mathbf{C}_2$ e $\mathbf{C}_4$, devido estes genes g mudarem de *cluster* quando mudamos o valor de β e consequentemente mudarem sua condição de com evidências para diferença para sem evidências para diferença, ou vice-versa. Ou seja, o modelo com mistura infinita é sensível a escolha dos hiperparâmetros das distribuições *a priori*, levando a diferentes conclusões, isto é, a formação de *cluster*s com quantidades diferentes de genes e consequentemente a detecção de diferentes genes g com evidências para níveis de expressão diferentes.

Acreditamos que este fato pode ser resolvido a partir da opinião de um especialista da área genética, nos fornecendo informação para definirmos os valores dos hiperparâmetros e assim trabalhar com informação *a priori* sobre as medidas d_g, $g = 1, 2, ..., G$, e obter *clusters* satisfatórios com relação as medidas na "fronteira" dos *clusters* formados.

Dessa forma, se temos informação para definirmos os valores dos hiperparâmetros, acreditamos que a utilização do modelo com mistura infinita e do critério DIC se mostra como ferramentas estatísticas adequada para identificar *clusters* de genes com e sem evidências para diferença baseados no efeito de tratamento, quando comparamos uma situação de tratamento com uma condição de controle.

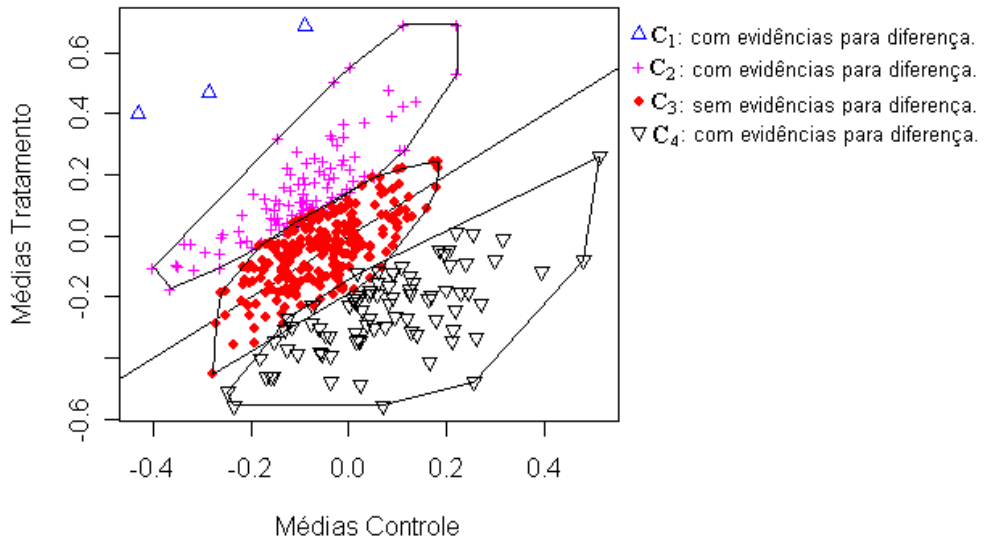

Figura 20: Médias controle e tratamento e *cluster*s formados, com $\beta = 2$.

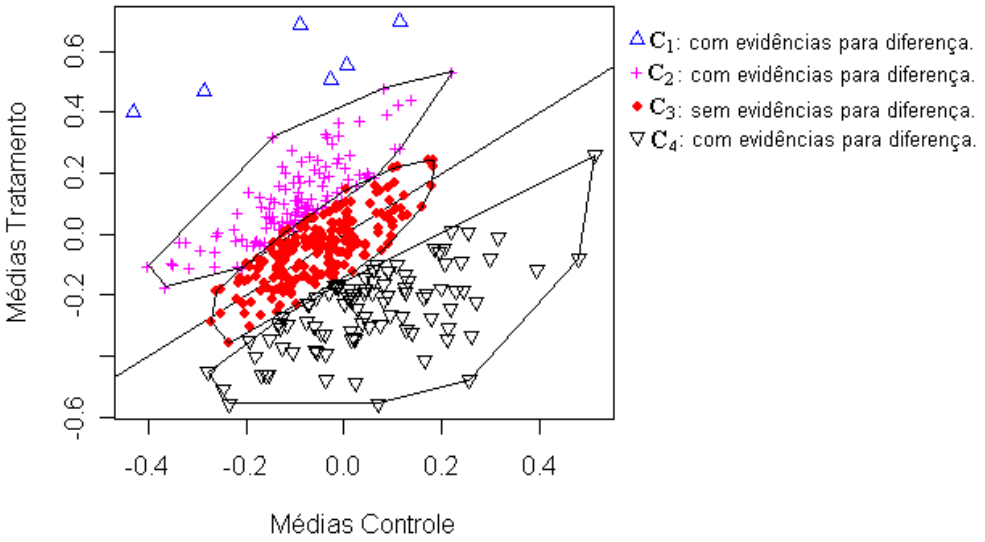

Figura 21: Médias controle e tratamento e *cluster*s formados, com $\beta = 1$.

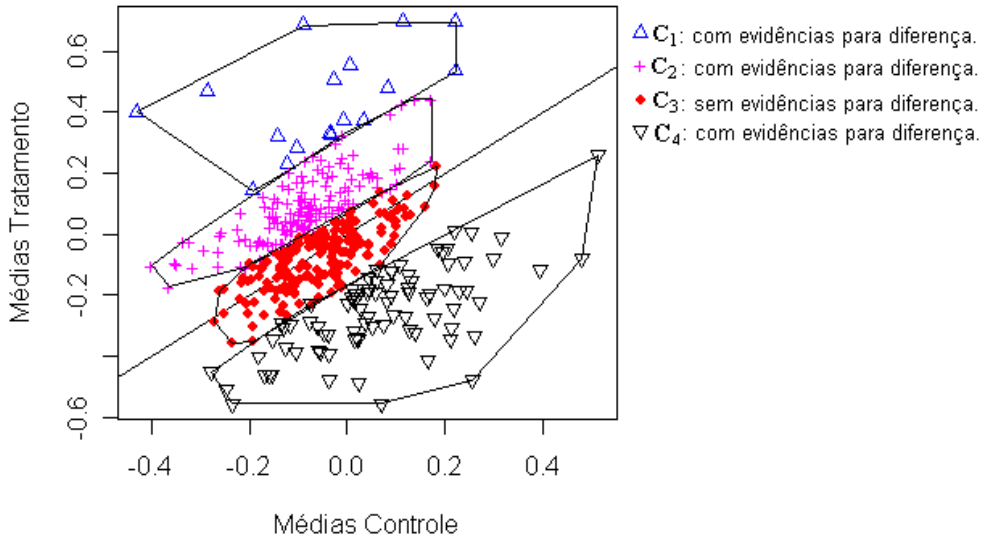

Figura 22: Médias controle e tratamento e *cluster*s fomados, com $\beta = 0,5$.

Capítulo 6

Considerações Finais

Nesta dissertação descrevemos, adaptamos e desenvolvemos métodos estatísticos aplicados à análise da expressão gênica, com objetivo de identificar genes que apresentam evidências para níveis de expressão diferentes quando comparamos situações de interesse (tratamentos) com uma situação de controle.

Os aspectos relacionados ao desenvolvimento dos experimentos com arranjos de DNA são abordados no Capítulo 1. Neste Capítulo descrevemos o processo de desenvolvimento para obtenção dos níveis de expressão dos genes, nas situações de tratamento e controle para posteriormente serem analisadas.

No Capítulo 2, descrevemos e aplicamos o teste t proposto por Baldi e Long (2001) na análise da expressão gênica através de um estudo de simulação e com a aplicação a dados de níveis de expressão gênica reais obtidos de um experimento realizado com as células da bactéria *Escherichia Coli* (ver Arfin *et al.*, 2000). Tanto na simulação quanto na aplicação o teste t mostrou uma deficiência, pois detectou evidências para diferença somente quando a diferença estava presente na média com as variâncias sendo próximas (ver, por exemplo, Figuras 2 e 3).

Motivados pelo trabalho de Baldi e Long (2001) propomos para análise da expressão gênica a utilização da inferência bayesiana e dos métodos de seleção de modelos fator de Bayes e DIC. Para a análise bayesiana utilizamos distribuições *a priori* conjugadas e pouco informativas. Para o cálculo do fator de Bayes utilizamos uma aproximação via método Monte Carlo.

Como o fator de Bayes é influenciado pelas distribuições *a priori*, isto é, quando com-

paramos dois modelos, um com distribuição *a priori* informativa e outro com distribuição *a priori* não informativa para os parâmetros, o primeiro modelo é privilegiado pelo fator de Bayes[1]. Devido a isto, utilizamos uma equalização para os parâmetros das distribuições *a priori* dos modelos M_0 e M_1 de modo que as distribuições *a priori* sejam equivalentes e dessa forma buscamos não privilegiar um dos modelos, M_0 ou M_1.

Tanto na simulação quanto na aplicação, o fator de Bayes e o critério DIC se mostraram sensíveis à escolha dos hiperparâmetros, ou seja, devemos ter certa atenção para definirmos os valores dos hiperparâmetros.

Com o objetivo de reduzir as restrições para as análises dos dados de níveis de expressão gênica, no Capítulo 4 descrevemos um dos principais métodos *a priori* em inferência bayesiana não paramétrica, conhecida como processo Dirichlet.

Com a utilização do processo Dirichlet (PD) *a priori* utilizamos à inferência bayesiana com uma especificação *a priori* não paramétrica sobre todas as classes de possíveis funções de distribuição $G(y)$ de uma variável Y, onde $G(y) = P(Y \leq y)$ (Ver, Ibrahim 1998). Assim, temos que se uma distribuição de probabilidades desconhecida G possui distribuição *a priori* definida por um processo Dirichlet, $G \sim PD$, significa que sua medida induzida g, representa uma medida de probabilidade pertencente a um espaço de funções de distribuição, portanto não podendo ser indexada por um número finito de parâmetros, o que caracteriza o processo Dirichlet como sendo não paramétrico (ver Walker *et al.*, 1999).

Descrevemos o modelo de misturas de processos Dirichlet (MPD) que é essencialmente um modelo bayesiano hierárquico que remove a suposição de uma distribuição paramétrica para os parâmetros do modelo paramétrico em estudo, substituindo-a por uma distribuição qualquer G que possui distribuição *a priori* definida por um processo Dirichlet com parâmetro de concentração M e distribuição base G_0 (isto é, medida base Mg_0).

Aplicamos o modelo de misturas de processo Dirichlet (MPD) na análise da expressão gênica utilizando modelos conjugados. Pois o uso de modelos conjugados torna a implementação computacional mais viável e a amostragem mais eficiente (ver MacEachern e Muler, 1998, Neal, 1998 e Dahl, 2002). Devido a isto, utilizamos os hiperparâmetros M, λ, α e β fixos e de tal forma que as distribuições *a priori* obtidas sejam não informativas

[1] Para maiores detalhes ver Kass e Raftery (1995) ou Missão (2004).

(ver Tabela 21 e Figura 15).

Desenvolvemos para o modelo MPD o mesmo estudo de simulação realizado para o teste t, para o fator de Bayes e para o critério DIC e aplicamos este aos dados reais obtidos do experimento realizado com as células da bactéria *Escherichia Coli*.

Como o fator de Bayes e o critério DIC, o modelo MPD se mostra sensível a escolha dos hiperparâmetros. Ou seja, alterando o valor dos hiperparâmetros os genes detectados com evidências para diferenças são diferentes. À medida que temos informação *a priori* o modelo MPD se mostra confiável na detecção dos genes com evidências para diferença.

Muitas vezes o interesse é de identificar e analisar grupos (ou *cluster*) de genes. Devido a isto, no Capítulo 5 descrevemos e desenvolvemos um procedimento de "aglomeração" de níveis de expressão gênica[2], baseados nos efeitos de tratamento, utilizando um modelo com mistura infinita que é equivalente ao modelo de misturas de processo Dirichlet.

Como no modelo de misturas de processos Dirichlet, utilizamos os hiperparâmetros fixos de modo a manter a conjugação e facilitar a implementação computacional e tornar a amostragem mais eficiente (ver Dahl, 2002).

O modelo com mistura infinita também se mostra sensível a escolha dos hiperparâmetros. Esta sensibilidade ocorre nos genes com estatísticas do efeito de tratamento d_g na "fronteira" dos *cluster*s formados, devido estes trocar de *cluster* quando mudamos o valor dos hiperparâmetros. Para os genes com medida d_g distantes da reta $y = x$, o método se mostra eficiente, mesmo para distribuições *a priori* não informativas, detectando estes genes como pertencentes a um *cluster* com evidências para diferença.

Com os resultados obtidos com procedimento de simulação e com os dados reais, obtidos do experimento realizado com as células da bactéria *Escherichia Coli* (ver Arfin *et al.*, 2000), utilizados nesta dissertação, temos que

- O teste t não se mostra como um método estatístico eficiente para se obter resultados satisfatórios na análise da expressão gênica. Pois detecta evidências para diferenças de médias com variâncias de tratamento e de controle razoavelmente estáveis, como pode ser observado pelos valores da Tabela 1 e pelas Figuras 2 e 3;

- Os resultados obtidos com o fator de Bayes e com o DIC são similares. Comparado ao teste t, estes dois métodos apresentam um desempenho melhor, pois detectam evidências

[2] que chamamos de *cluster*.

para diferença tanto com relação à variação na média quanto com relação à variação na variância, ou ambos (ver Tabelas 4, 5, 6 e 9, 10, 11 e Figuras 4 a 9);

- O modelo de misturas de processos Dirichlet (MPD) detecta evidências para diferença de médias, independente se esta possui ou não diferença de variâncias entre as observações de tratamento e as observações de controle. Comparado ao teste t, o modelo MPD apresenta um desempenho melhor. Pois, por exemplo, na aplicação este método detectou os genes mais distantes da reta $y = x$, no gráfico das médias, com evidências para diferença enquanto que o teste t não os detectou (ver Figuras 2 e 3 e Figuras 12, 13 e 14). Grande parte dos genes identificados com evidências pelo modelo MPD também foram identificados pelo fator de Bayes e pelo DIC, como pode ser observado pelas Figuras 4 a 9 e Figuras 12, 13 e 14;

- Com a utilização do modelo com mistura infinita e do DIC identificamos *clusters* de genes com e sem evidências para diferença, devido ao efeito de tratamento. Na aplicação os genes identificados com evidências para diferença pelo teste t e pelo modelo MPD também foram identificados por este método. Com relação ao identificados com evidências para diferença pelo fator de Bayes e pelo DIC, somente os genes com média de tratamento e de controle próximos a reta $y = x$ no gráfico das médias, identificados pelo fator de Bayes e pelo DIC (ver Figuras 4 a 9), não foram identificados pelo modelo com mistura infinita e pelo DIC (ver Figuras 20, 21 e 22).

Enfim, esta dissertação aborda cinco métodos estatísticos que pode se aplicados na análise da expressão gênica: teste t, fator de Bayes, DIC, modelo MPD e modelo com mistura infinita e DIC. Cada um com suas características, como discutidos no texto.

Referências Bibliográficas

[1] Aitkin, M. (1991) Posterior Bayes Factor (with discussion). *Journal of the Royal Statistical Society*, **53**, 111-142.

[2] Alston, C., Kunert, P., Low choy, S., McVinish, R., Mengersen, K. (2005) Bayesian model comparison: Review and Discussion. http://www.stat.auckland.ac.nz/~iase/publications/13/Alston-Kuhnert-Low_Choy McVinish-Mengersen.pdf.

[3] Antoniak, C. E. (1974) Mixture of Processes With Applications to Bayesian Non-parametric Problems. *The Annals of Statistics*, **2**, 1152-1174.

[4] Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S. and Hatfield, G. W. (2000) Global Gene Expression Profiling in Escherichia Coli K12. *J. Biol. Chem.*, **275**, 29672-29684.

[5] Baggerly, K. A., Coobes, K. R., Hess. K. R., Stivers, D. N., Abruzzo, L. V. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *Jornal Computational Biology*, **8**, 639-659.

[6] Baldi, P. and Long, D. A. (2001) A Bayesian Framework for the Analysis of Microarray Expression Data: regularized t-test and statistical inferences of gene changes. *Bioinformatcs*, **17**, 509-519.

[7] Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999) Clustering gene expression patterns. *Jornal Computational Biology*, **6**, 281-297.

[8] Best, N., Cowles, M. k., Vines, K. CODA - Convergence Diagnosis and Output Analysisi Software for Gibbs Samplig Output. Version 0.4. Biostatistics Unit MRC, Cambridge, Inglaterra, 1995 (Relatório técnico).

[9] Blackwell, D. and MacQueen, J.B. (1973) Ferguson Distribution Via Pólya Urn Schemes. *The annals of Statistics*, **1**, 353-355.

[10] Bush, C. A., and MacEachern, S. N. (1996) A semi-parametric Bayesian model for randomised block designs. Biometrika, **83**, 275-286.

[11] Carneiro, N. P., Carneiro, A. A., Guimarães, C. T. e Paiva, E. (2001) Desvendando o Código Genético. *Biotecnólogia Ciência e Desenvolvimente*, 50-58.

[12] Celeux, G., Hurn, M., and Robert, C. P. (2000) Computational and Inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95** (3), 957-979.

[13] Celeux, G., Forbes, F., Robert, C. P. e Ttterington, D. M. (2005) Deviance Information Criteria for Missing Data Models. http://www.ceremade.dauphine.fr/~xian/cfrt03.pdf.

[14] Dahl, D. B. (2002) Modeling Diferential Gene Expression Using a Dirichlet Process Mixture Model. http://www.stat.tamu.edu/~dahl/em4ged/jsm2003paper.pdf.

[15] Dempster, A. P. (1974) The Direct use of likelihood for significance testing. In proceeding of conference on Foundation Question in Statistical Inferene. 335-352. Department of theoritical Statistics, university of Aarhus.

[16] DeRisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. and Trent, J. M. (1996), Use of a cDNA Microarray to Analyze Gene Expression Patterns in Humam Cancer. *Nature Genetics*, **14**, 457-460.

[17] Diebolt, J., Robert, C. P. (1994) Estimation of finite mixture distribuition thoug bayesian sampling. *Journal of the Royal Statistical Society*, B, **56**, 2, 363-375.

[18] Do, K.A; Müller, P. Tang, F.(2002) A Bayesian Mixture for Differential Gene Expression. http://odin.mdacc.tmc.edu/~pm/pap/DMT02.pdf.

[19] Dudoit, S., Yang, Y., Callow, M., and Spped, T. (2000) Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Technical Report, University of California, Berkeley, Dpt. Statistics.

[20] Efron, B., Tibishirani, R., Storey, J. D., and Tusher V. (2001) Empirical Bayes Analysis of a Microarray Experiment. *Journal american statistical association*, **96**, 1151-1160.

[21] Efron, B., Tibishirani, R., Goss, V., and Chu, G. (2000) Microarrays and their use in a Comparative Experiments. Stanford Technical Report 213.

[22] Escobar, M. D. (1994) Estimating Normal Means With a Dirichlet Process Prior. *Journal of the American Statistical Association*, **89**, 268-277.

[23] Escobar, M. D., and West, M. (1995) Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**, 577-588.

[24] Felix, J. M., Drummond, R. D.; Nogueira, F. T. S.; Junior, V. E. R.; Jorge, R. A.; Arruda, P.; Menossi, M. (2002) Genoma Funcional.*Biotecnólogia Ciência e desenvolvimento*, **24**, 60-67.

[25] Ferguson, S. T. (1973) A Bayesian Analysis of Some Nonparametric Problems. *The annals of statistcs*, **2**, 209-230.

[26] Gelfand, A. E., Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *Journal American Statistics Association*, **87**, 523-531.

[27] Gelfand, A. E., and Kottas, A. (2002) A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **11**, 289-305.

[28] Gelmam, A. B., Carlim, J. S., Stern, H. S.; Rubin, D.B. (1995) *Bayesian Data Analysis*. , London, New York Washington; Chapmam Hall

[29] Gilks, W. R., Richardson, S. and Spiegelhalter (1996) *Markov Chain Mont Carlo in Pratice*. London New York Washington; D.C. Chapmam Hall.

[30] Gosh, D. and Chinnaiyan, A. M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275-286.

[31] Guyton, A. C. (1998) *Fisiologia Humana*. Rio de Janeiro. Guanabar Koogan S.A.

[32] Hastings, W. R. (1970) Monte Carlo Samling methods using Markov chain and their applications. *Biometrika*, **57**, 97-109.

[33] Ibrahim, J., G., Chen, M. H., Sinha, D. Y., (2001) *Bayesian Survival Analysis*. Spinger.

[34] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Spped, T. (2003) Explortion, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264.

[35] Jain, S., and Neal, R. M. (2000) A Split-merge markov chain monte carlo procedure for the Dirichlet process mixture model. Technical Report, Department of Statistics, University of Toronto.

[36] Kamb, A., Ramaswami, M. (2001) A Simple Method For Statistical Analysis of Intensity Difference in Microarray-Derived Gene Expression Data. *BMC Biotechnology*, 1-8.

[37] Kass, R., and Raftery, A. (1995) Bayes Factor. *Journal of the American Statistical Association*, **90**, 773-795.

[38] Kraft, C. H. (1964) A Class of distribution function process which have derivaes. Jornal probability, **1**, 385-388.

[39] Kraft, C. H., van Eeden, C. (1964) Bayesian bio-assay. *The annals of matematics of the statistc*, **35**, 886-890.

[40] Leite, J. G. e Singer, J. M. (1990) *Métodos assintóticos em estatística: Fundamentos e aplicações*. São Paulo, Associação brasileira de estatística (ABE).

[41] MacEachern, S. N. (1994) Estimating Normal Means With a Conjugate Style Dirichlet Process Prior. *Comunication in Statistics: Simulation an computation*, **23**, 727-741.

[42] MacEachern, S. N., and Müller, P. (1998) Estimating Mixture of Dirichlet Process Models. *Journal of Computational an Graphical Statistics*, **7**, 223-238.

[43] McLachlan, G. J., Bean, R. W., and Peel, D. (2002) A Mixture Model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422.

[44] Medvedovic, M. and Sivaganesan, S. (2002) Bayesian Infinite Mixture Model Based *cluster*ing of Gene Expression Profiles. *Bioinformatics*, **18**, 1194-1206.

[45] Missão, E. C. M. (2004) *Uma revisão do fator de Bayes com aplicação à modelos com misturas.* Dissertação de Mestrado, DEs-UFScar.

[46] Neal , R. M. (1998) Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Technical Report 4915. Department of Statistics, university of Toronto. http://cs.toronto.edu/~redford/mixmc.abstract.html.

[47] Newton, M., Kendziorski, C., Richmond, C., Blatter, F., and Tsui, K. (2001) On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data. *Journal of Computational Biology*, **8**, 37-52.

[48] Pan, W., Lin, J., and Le, C. T. (2002) Model-based cluster analusis of microarray gene expression data. *Genome Biology*, **3**, 1-8.

[49] Passos, G. A. S., Jordan, C. N. B. (2001) Projeto Transcriptoma. *Biotecnólogia Ciência e Desenvolvimento*, 34-37.

[50] Paulino, C. D., Turkman A. A., Murteira B. (2003) *Estatística Bayesiana*. Edição da Fundação Calouste Gulbenkian.

[51] Rasmusses, C. F. (2000) The infinite Gaussian Mixture Model. http://bayes.imm.dtu.dk

[52] Richardson, S., and Green, P. (1997) On Bayesian analysis of mixture with an unknow number of components (with discussion). *Journal of the Royal Statistical Society*, B, **64**(3), 583-639.

[53] Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995), Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

[54] Schena, M., Shalon, D., Heller, R.., Chai, A., Brown, P. O. and Davis, R. W. (1996) Parallel humam genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceeding of the National Academy of Sciences*, **93**, 10614-10619.

[55] Sethuramam, J. (1994) A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639-650.

[56] Spiegelhalter, D. J., Best, N. G., Carlim, B. P., van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, B, **64** (3) 583-639.

[57] Stephens, M. (2002) Bayesian analysis of mixture models with an unknow number of componenets - an alternative to reversible jump methods. *The Annals of Statistics*, **28**, 40-74.

[58] Venables, W. N.; Ripley, B. D. (1998) *Modern Applied Statistics With S-Plus*, *Statistics e computing*. Springer.

[59] Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999) Bayesian Nonparametric Inference For Random Distributions and Related Functions. *Journal of the Royal Statistical Society*, 1369-7412.

[60] West, M., Müller, P., Escobar, M. D. (1994) Hierarquical priors and mixture models with applications in regression and density estimation. In P. R. Freemam and A. F. M. Smith (editors), Aspects of Uncertainty, 363-386. John Wiley.

[61] Wu, T. D. (2001) Analyzing gene expression data from DNA microarray to identify candidates genes. *Jornal of Pathology*, **195** (1), 53-65.

[62] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987.